2010

AUDIO SCENE SEGEMENTATION USING A MICROPHONE ARRAY AND AUDITORY FEATURES

Harikrishnan Unnikrishnan
University of Kentucky, harikrishnan@uky.edu

ABSTRACT OF THESIS

AUDIO SCENE SEGEMENTATION USING A MICROPHONE ARRAY AND AUDITORY FEATURES

Auditory stream denotes the abstract effect a source creates in the mind of the listener. An auditory scene consists of many streams, which the listener uses to analyze and understand the environment. Computer analyses that attempt to mimic human analysis of a scene must first perform Audio Scene Segmentation (ASS). ASS find applications in surveillance, automatic speech recognition and human computer interfaces. Microphone arrays can be employed for extracting streams corresponding to spatially separated sources. However, when a source moves to a new location during a period of silence, such a system loses track of the source. This results in multiple spatially localized streams for the same source. This thesis proposes to identify local streams associated with the same source using auditory features extracted from the beamformed signal. ASS using the spatial cues is first performed. Then auditory features are extracted and segments are linked together based on similarity of the feature vector. An experiment was carried out with two simultaneous speakers. A classifier is used to classify the localized streams as belonging to one speaker or the other. The best performance was achieved when pitch appended with Gammatone Frequency Cepstral Coefficeints (GFCC) was used as the feature vector. An accuracy of 96.2% was achieved.

KEYWORDS: Audio Scene Segmentation, Sound Source Tracking, Computational Auditory Scene Analysis, Microphone Arrays, Speaker Recognition.

Harikrishnan Unnikrishnan

11/23/2009

AUDIO SCENE SEGEMENTATION USING A MICROPHONE ARRAY AND AUDITORY FEATURES

By

Harikrishnan Unnikrishnan

Kevin D. Donohue
Director of Thesis

Stephen D. Gedney
Director of Graduate Studies

11/23/2009

RULES FOR THE USE OF THESES

Unpublished theses submitted for the Master's degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgments.

Extensive copying or publication of the thesis in whole or in part also requires the consent of the Dean of the Graduate School of the University of Kentucky.

A library that borrows this thesis for use by its patrons is expected to secure the signature of each user.

Name	Date

THESIS

Harikrishnan Unnikrishnan

The Graduate School

University of Kentucky

2009

AUDIO SCENE SEGEMENTATION USING A MICROPHONE ARRAY AND AUDITORY FEATURES

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in the
College of Engineering
at the University of Kentucky

By

Harikrishnan Unnikrishnan

Lexington, Kentucky

Director: Dr. Kevin D. Donohue, Professor of Electrical Engineering

Lexington, Kentucky

2009

Dedicated to all the teachers to whom I owe everything I know.

Acknowledgements

I express my sincere gratitude to my advisor Dr. Kevin D. Donohue for not only guiding me through my research but also for giving me confidence and space to think independently.

I would like to thank the committee member Dr.-Ing Jens Hannemann, Dr. Laurence G. Hassebrook and Dr. Sen-ching Cheung for their valuable time and insight. I thank Satoru Tagawa, Phil Townsend, and other colleagues at Center for Visualization and Virtual Environments whose cooperation made my time in the lab so edifying.

I appreciate the support I received from my family and friends. The encouragement given by my parents Dr. Unnikrishnan K. and Mrs. Radha Unnikrishnan gave me the energy to acclimatize socially and culturally to a new place and still not lose focus on the task of completion of my degree. I thank all the friends who made my stay at the university a memorable one.

Table of Contents

List of Tables

List of Figures

List of Files

ETD_AudioSceneSegmentationThesis.pdf

Chapter 1. Introduction

Auditory scene analysis (ASA) is the perceptual process by which a listener make sense of the auditory world consisting of multiple sources. The composite signal that enters our ears is used for the purpose. The human auditory system consisting of ear canals, ear-drums, cochlea and auditory nerves produce nerve impulses. These impulses are received by the brain, which uses it along with prior knowledge, redundancy in speech and linguistic considerations (grammar and semantics) to identify distinct objects/events. The objective of Computational Auditory Scene Analysis (CASA) is to make a computational model of ASA. CASA almost always assumes that no *a priori* knowledge of source locations or number of sources is available.

1.1. Terms Related to CASA

The following are the important terms used in CASA.

Acoustic source –"the concrete, physical manifestation of a sound wave" [1]. The source can be a human speaker, music being played or a car driving pass a listener etc.

Cues - Cues in the context of ASA are the features which represent all or part of a sound. They are the means by which a certain goal is achieved in ASA. The goal can be listening to a particular source in a backdrop of noises or other interfering sources, or it can be just identifying the location of a speaker. Some cues that are used are pitch, onset, offset, amplitude modulation or envelope and spatial location[2].

Tracks – They are the outcomes of low level feature extraction. They are formed by linking continuous points of the sound signal in a time-frequency(TF) space. The principle of proximity in time and frequency[1] serves as the basis for defining a track. Figure 1.1 shows the spectrogram of an utterance of the vowel sound 'aa'. Each pixel defines a point in TF space and the color represents its intensity. The continuous pixels which can be grouped together on the basis of their intensity form a track.

Segments – Segments are formed by combining the tracks or regions in TF space which are related. If the source to be segregated is periodic in nature, the harmonically related tracks are grouped to form segments. In Figure 1.1 the distinct tracks are grouped together as they are related harmonically. In case of unvoiced signal, cues like onset and offset are used[3]. Computationally auditory segmentation is analogous to image segmentation. Binary gain masks and region growing (cluster analysis) are used in this stage[3].

Auditory snapshot - Consider the case of image segmentation, where non-overlapping segments are identified. Segments consist of a group of pixels which represent one object. The union of segments define the image. Similarly auditory segments existing at a given instant of time define the space of interest; analogous to an image. It can be called as an *Auditory Snapshot (AS)***.**

Auditory Stream - Auditory stream denotes the abstract, conceptual effect a source has in the mind of the listener[1]. An auditory stream is always associated with a source. In a computational model streams are obtained by linking segments across time. This is an application specific task and is the most open problem in CASA.

Scene – Scene is a continuous series of AS which are linked by a high percentage of streams. A scene change is characterized by a change in dominant sources. Two scenes cannot exist at the same time. The task of segmenting the scene into streams can be termed as Audio Scene Segmentation (ASS).

Event – An event is a physical happening which corresponds to termination of one scene and the beginning of another. The event is marked by an appreciable change in the state (amplitude, pitch, location, etc.) of multiple sources and/or the introduction or termination of sources. An example for an event could be musician starting or finishing his performance and the performance constitute the scene.

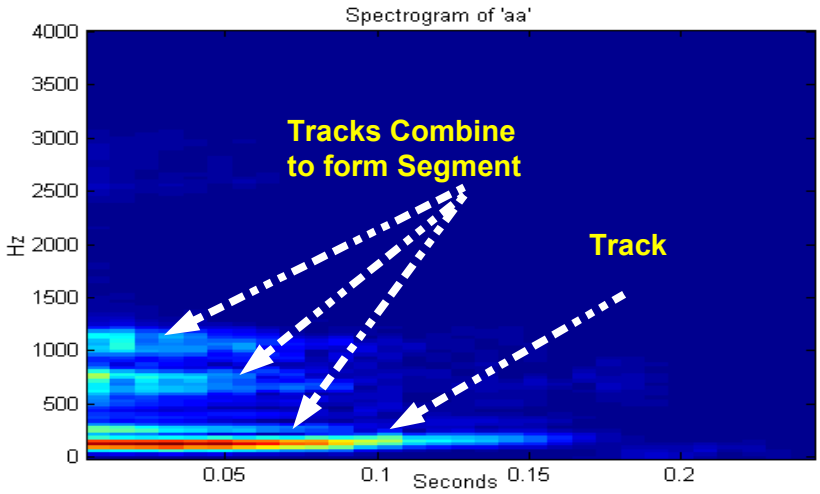

Figure 1.1: Spectrogram of 'aa'. Figure depicts track, segments

1.2. **Principle Stages of CASA**

A CASA system is depicted in Figure1.2. The various stages involved are[4] :

Peripheral processing: It is the process of making a time-frequency representation of the audio signal.

Low level feature extraction: The tracks in time-frequency space are extracted at this stage. Only one frequency will be associated with a track at any given instant of time. Tracks are continuous in nature.

Mid-level grouping: The tracks are grouped together to form the building blocks for high level grouping. The tracks in the same group may be harmonically related or have similar contours in time-frequency space.

High level streaming: This is the process of linking segments across time to form a stream. A representation of an object is formed in this stage.

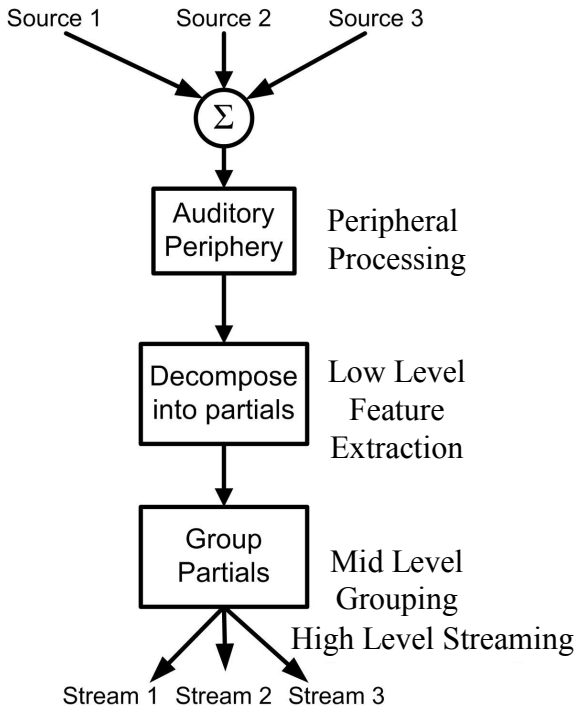

Figure 1.2: Stages Involved in CASA.
Adapted from[4].

1.3. **Computational Auditory Scene Analysis and Acoustic Scene Analysis (AcSA)**

Wang and Brown [3] propose a more specific definition for CASA

"... It (CASA) is the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene ...".

This definition makes CASA applicable for fields like developing hearing aids. Though CASA does not in any way restrict itself to the modeling of psychoacoustic system of humans, many works in this field is based on it[3].

AcSA is defined by [5] as

"...the task of extracting information contained in the acoustic wave-field, such as the waveform itself or parameter describing the source of the wave-field...."

AcSA relies on classic signal processing algorithms. A wave-field produced by a source is spread spatially and in time. Hence the use of microphone arrays is a standard way of completing the task. This differs from CASA in that the modeling of human hearing system is not directly used in developing its methods; however the outcomes may be similar.

The technique developed for CASA and AcSA can be combined for improving the performance of Audio Scene Segmentation. Microphone arrays can provide spatial location with greater accuracy than the human auditory system. The information provided by acoustic waveform modeling along with information from human ASA model can provide superior performance to that of techniques developed using conventional signal processing tools.

1.4. **Objective**

The goal is to extract streams with enhanced intelligibility using all the available methods. Each stream is potentially an input to an automatic speech recognition (ASR) system. Also events can be used to trigger automated processes. Such a system would also find application in simultaneous sound source tracking. Steered Response Coherent Power (SRCP) estimated using a microphone array has been used for the estimation of source locations in previous works[6][7]. The task was to detect the presence of sound sources at any location within a field of view. This thesis aims to link the detected sound sources across time frames. If sound sources are within a time and distance threshold of each other, this can be used to link detected sources together over time. But intervals of silence in which the source moves to a different location complicate this process and additional cues are needed to link sources to the same object/person when separated by periods of silence.

This thesis uses high level features used in CASA and Automatic Speaker Recognition systems along with the spatial location to link the same source across the time frames. The system would find applications in multiple sound source tracking and advanced human computer interfaces[8]. This would allow the system to focus the attention on one

speaker of interest among multiple sources irrespective of the location within the field of view and the state of motion.

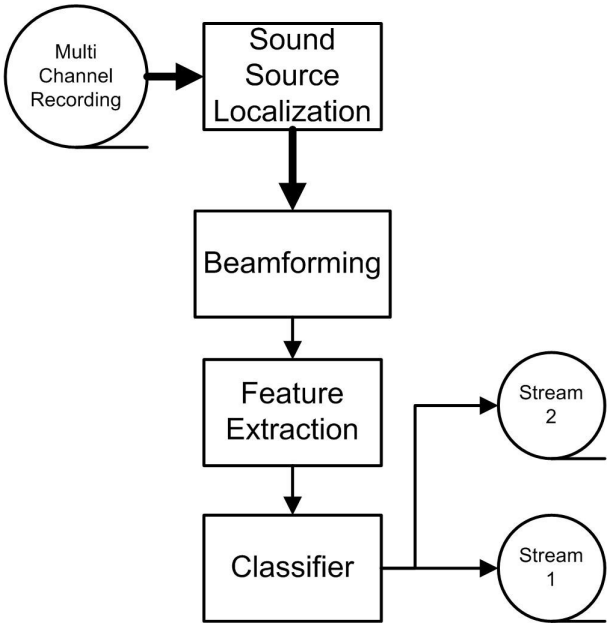

Figure 1.3: Functional block diagram of the ASS system.
It takes the multi channel recording of FOV as the input and gives the streams associated with each source as the output. Thick lines represent multi channel data.

1.5. **Hypothesis**

This thesis proposes that streams associated with each source can be extracted from a multi channel recording using sound source localization, beamforming and CASA feature analysis in sequence. The proposed system is depicted in Figure 1.3. The focus is on forming the streams first with spatial and temporal proximity and then linking these local streams using CASA and speaker recognition features.

1.6. **Approach**

The objects considered in all auditory scenes for this work are human speakers. The approach examined in this thesis uses several levels of detection and classification to establish a relationship between speakers in a scene over time and space. At the lowest level, the location of a speaker is detected in each time frame applying sound source

localization techniques[6] with microphone arrays. On the second level detected sources are grouped together based on space and time proximity. If there is limited silence between two consecutive detections, proximity in space could be used to identify the streams linked to each source. In cases where source is at $(x_{1,} y_{1,} z_1)$ and remains silent for some time and then is again detected at another location $(x_{2,} y_{2,} z_2)$ spatial coordinates cannot be used. Hence a third level is required that uses others features that remain relatively invariant for each speaker to link detect segments together. Once a source is detected, a delay and sum beamformer is used to enhance the source signal before feature extraction. The features listed below are considered in this work. These are analyzed for its effectiveness in ASS, and include:

1. Pitch
2. Envelope energy (loudness)
3. Rate of change in speech.
4. Mel Frequency Cepstrum Coefficients (MFCC) and its first order delta.
5. Gamma-tone Feature Cepstral Coefficients (GFCC) and its first order delta.
6. Vocal chord transfer function
7. Center of mass of vocal chord transfer function.

1.7. **Outline**

A review on Beamforming, specifically Delay and Sum Beamforming(DSB) is presented in Chapter 2. Then Sound Source Localization(SSL) using SRCP and the lowest level of processing is explained in Chapter 3. In Chapter 4, the identified sources are linked using proximity in spatial location. The result of Auditory Scene Segmentation using spatial cues is shown and the need for using auditory cues is explained. Chapter 5 introduces and analyzes the auditory features which may be used for Auditory Scene Segmentation. The better performing feature is identified here. In Chapter 6 a classifier which uses the suitable features to perform Scene Segmentation in the case of two simultaneous speakers is introduced. The work is summarized in Chapter 7.

Chapter 2. Beamforming

2.1. Introduction

Beamforming is the process of enhancing the target signal contaminated by interfering sources and ambient noises by spatial filtering[9][10][11]. An array of sensors, (microphones in case of audio) is employed for this. If the source signal and interfering signal originate from different spatial locations, beamforming can be used to enhance the desired signal. The simplest form of the beamformer is the Delay and Sum Beamformer(DSB). DS beamformer and its design issues are discussed. Beamfoming is used in this thesis for steering the array to focus its attention to a point in Field Of View(FOV) for Sound Source Localization(SSL). After SSL, it is again used for source signal enhancement.

2.2. Delay and Sum Beamformer

A DS Beamformer[9][10] consists of basically two steps; delaying the signals received at each microphone array element by Time Difference Of Arrival (TDOA) and then adding up the delayed signals to obtain DSB output. Figure 2.1 depicts a DSB.

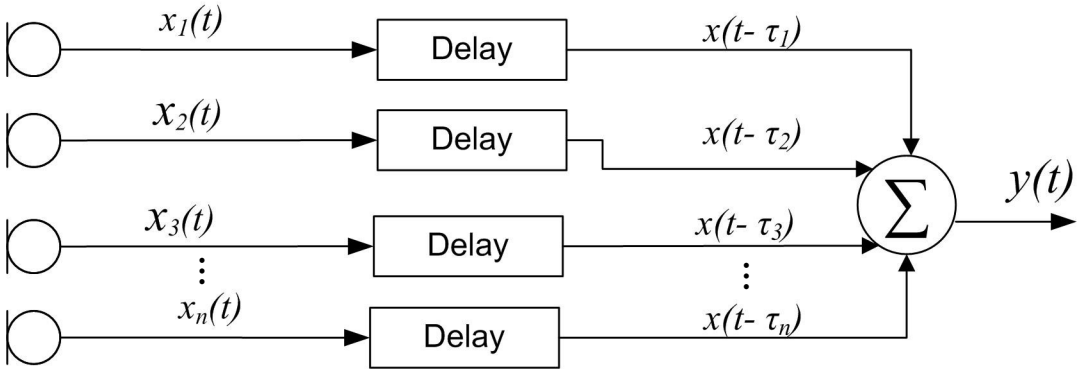

Figure 2.1: Delay and Sum Beamformer

In Figure 2.1 $x_n(t)$ denotes the signal at the n^{th} microphone and can be modeled as :

$$\begin{aligned} x_n(t) &= h(\Delta t, \vec{r}_s, \vec{r}_n) * \vec{s}(t - \Delta t_n) \\ &\quad + h(\Delta t, \vec{r}_i, \vec{r}_n) * b_i(t - \Delta t_{i,n}) + \eta_n(t) \end{aligned} \tag{2.1}$$

$\vec{r}_s$ is the target location and $\vec{r}_n$ are the microphone locations. $n = 1,2,3 \ldots, N$. N is the number of microphones. $s(t)$ is the direct path source signal, b_i are the first i significant reverberations of $s(t)$. $\vec{r}_i$ are locations from where i reverberations originate. Δt is the propagation delay of the sound from the source to the microphone. $\eta_n(t)$ is the additive uncorrelated noise. The sound sources other than the target also contribute to the uncorrelated noise. Let microphone 1 be the reference microphone which implies :

$$\Delta t_n = \tau_{ref} - \tau_n \quad ; \tau_1 = 0 \tag{2.2}$$

τ_n is the Time Difference Of Arrival between the n^{th} microphone and the 1st microphone. Substituting for Δt_n in Eq.2.1:

$$\begin{aligned} x_n(t) &= h(\Delta t, \vec{r}_s, \vec{r}_n) * s(t - \tau_{ref} + \tau_n) \\ &\quad + h(\Delta t, \vec{r}_i, \vec{r}_n) * b_i(t - \tau_{ref} + \tau_{i,n}) + \eta_n(t - \tau_{ref} + \tau_n) \end{aligned} \tag{2.3}$$

In order to beamform to any point in the FOV the unique combination $\tau_n s$ corresponding to that point is used. The DSB output is given by :

$$\begin{aligned} y(t) &= \sum_{n=1}^{N} x_n(t - \tau_n) \\ &= \sum_{n=1}^{N} h(\Delta t, \vec{r}_s, \vec{r}_n) * s(t - \tau_{ref}) + \sum_{n=1}^{N} h(\Delta t, \vec{r}_i, \vec{r}_n) * b_i(t - \tau_{ref} + \tau_{i,n} - \tau_n) + \sum_{n=1}^{N} \eta_n(t - \tau_{ref}) \end{aligned} \tag{2.4}$$

In Eq.2.4 while the uncorrelated noises are reduced by incoherent summatio, the effect of reverberations(correlated noise) cannot be completely reduced as the speech signals are strongly correlated with itself, especially for small lag $(\tau_n - \tau_{i,n})$ of the order of 20 – 40 ms.

2.3. **Directivity pattern and Design issues of DSB**

The response of the beamformer to different spatial locations of the target is known as its directivity/spatial pattern. The directivity pattern is dependent on the actual geometry of the microphone array. Consider a uniformly spaced linear microphone array(Figure 2.2.).

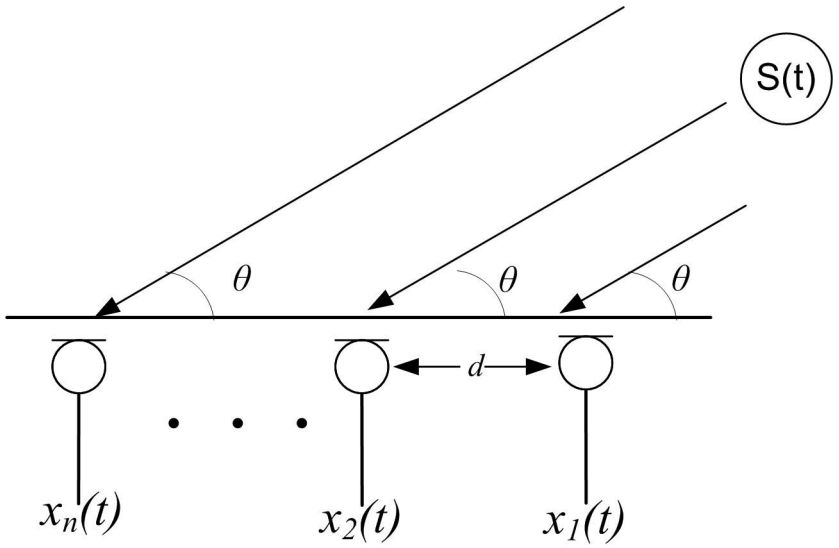

Figure 2.2: Equi-spaced linear microphone array
Source is located in the far field. θ is the angle of incidence.

TDOA for the n^{th} microphone is given by :

$$\begin{aligned} \tau_n &= (n-1)\alpha; \\ \alpha &= d\cos(\theta)/c \end{aligned} \tag{2.5}$$

where α is the TDOA between the second and first microphone. The Array response can be obtained by substituting unit impulse $\delta(t)$ for $x_n(t)$ in Eq.2.4 [9]. Also τ_n is substituted using Eq. 2.5.

$$y(t)=\sum_{n=1}^{N} \delta\left(t-(n-1)d\cos(\theta)/c\right) \tag{2.6}$$

Taking spatial (with respect to t and then θ) Fourier Transform of Eq.2.6 :

$$Y(\psi,\theta)=\frac{1}{N}\sum_{n=1}^{N}\left[\exp\left(j2\pi(n-1)f\,d/c\cos(\theta)\right)\right]\exp\left(-j2\pi(n-1)f\,d/c\left(\cos(\psi)\right)\right)$$
$$=\frac{1}{N}\sum_{n=1}^{N}\left[\exp\left(-j2\pi(n-1)f\,d/c\cos(\psi-\theta)\right)\right] \tag{2.7}$$

where $0\le\psi\le\pi$ is the directional angle and f is the frequency of the source. The magnitude response is then given by [9]:

$$A(\psi,\theta)=|Y(\psi,\theta)|$$
$$=\left|\frac{\sin\left[N\pi\,fd\left(\cos(\psi)-\cos(\theta)\right)/c\right]}{N\sin\left[\pi\,fd\left(\cos(\psi)-\cos(\theta)\right)/c\right]}\right| \tag{2.8}$$

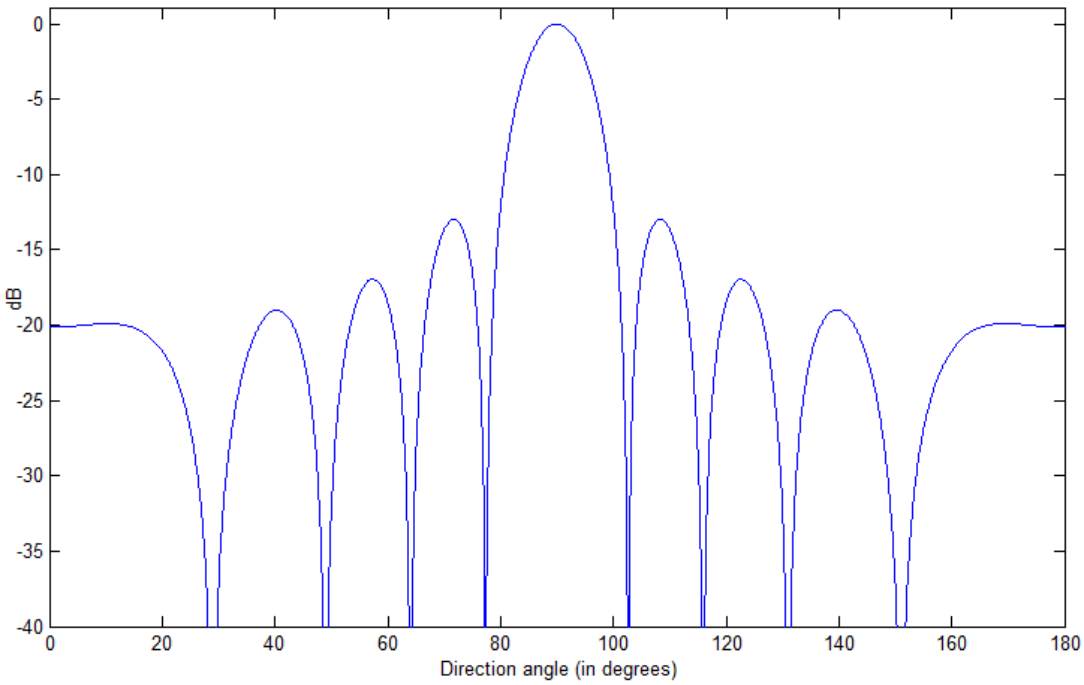

Figure 2.3: Beam Pattern of equi-spaced linear array
$\theta=90^{o}$, d = 8cm, c= 350m/s, f = 2000 Hz, N = 10. Beam pattern plotted using Eq. 2.8

It can be seen that as the inter microphone spacing d increases beam-width decreases. i.e. the directivity of the array improves. But an increase in d also causes an increase in side lobe intensity. Also if d is greater than $\lambda/2$ where $\lambda=c/f$ spatial aliasing occurs.

Figure2.4 Shows a case of spatial aliasing. It can be noted that there are two more side lobes with intensity at 0dB.

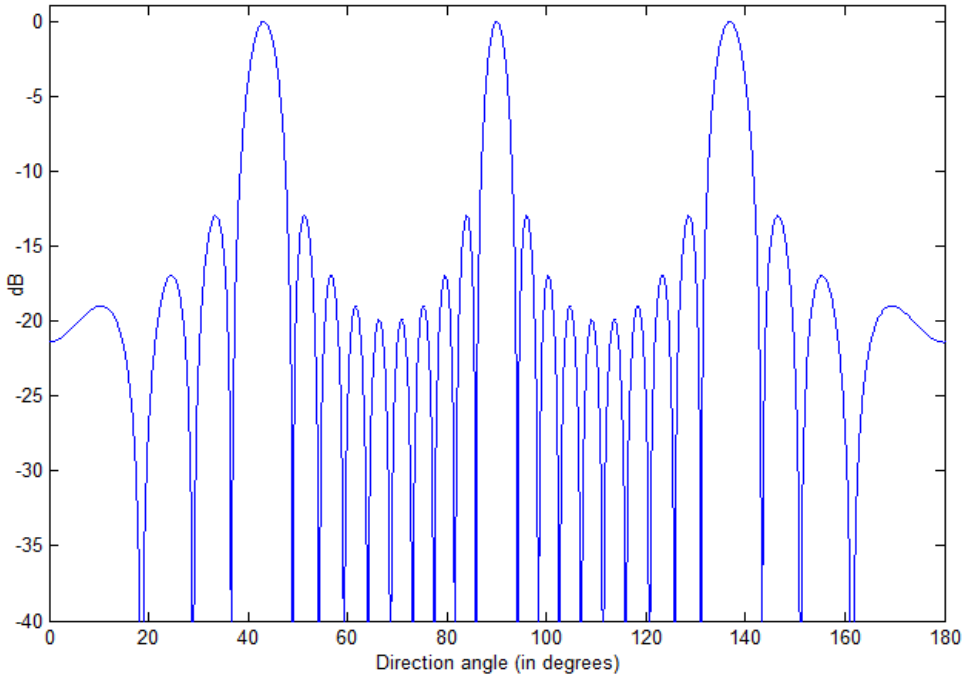

Figure 2.4: Beam Pattern of an equi - spaced linear array with spatial aliasing

d = 24cm, c= 350m/s, f = 2 kHz, N = 10. Beam pattern plotted using Eq.2.8 λ =17.5 cm

2.4. **Conclusion**

The Delay and Sum beamformer is the simplest type of beamformer. The target signal is enhanced by coherent addition. The uncorrelated noise signals tends to cancel each other by incoherent addition. Beam pattern and design issues of DSB were discussed for an equi-spaced linear microphone array. There exists a trade off between beam-width and side-lobe intensity. The design objective is to make the beam-width as narrow as possible while keeping side-lobe intensities at acceptable level. Spatial aliasing should be avoided and hence half the shortest wavelength in the input signal acts as the upper bound for the inter microphone spacing. A Simple DS beamformer is used in this work as a part of the SSL algorithm.

Chapter 3. Sound Source Localization

3.1. Introduction

The movement of a speaker will be localized in space for a given duration. If the location of a speaker is known at a given instance of time, the locus of points where he/she is present at any given time can be defined by the points within a circle(2D) or a sphere(3D) with current location as the center. The radius is a function of the maximum velocity with which he/she can move.

This chapter aims at estimating the spatial coordinates of the speakers present in the Field of View(FOV). Most popular Sound Source Localization(SSL) methods are based on Time Delay of Arrival(TDOA), Steered Response Power(SRP) or signal and noise subspace-based approaches[5]. All of them come under the domain of AcSA . TDOA based algorithms can be used only to locate a single source whereas SRP algorithms can be used in the scenarios where there are multiple sources. The SRP based approach used in this thesis is explained in detail in this chapter.

SSL is performed in overlapping windows of time to obtain a sequence of AS. The detections present in a sequence of ASs are linked together to achieve streaming. ASS by which streams are obtained are explained in the coming chapters (4 and 5).

3.2. SSL by Steered Response Power

In this approach a microphone array is made to beamform on each point in the FOV . The beamformer output signal power is then calculated. If it is above a predetermined/ adaptive threshold a source is deemed to be present.

The DS beamformer discussed in section 2.2 is used for steering the microphone array to each grid point. Let $I(x,y)$ be the grid points within the FOV. $I(x,y)$ can be defined by its distance from at least three non collinear reference points. The coordinates of microphone array elements acts as the reference points. If the speed of sound c is known or estimated,

the time taken for the sound to travel from $I(x,y)$ to the n^{th} microphone at $\vec{r}(x_n, y_n)$ is given by:

$$\Delta t_n = \frac{\sqrt{(x_n - x)^2 + (y_n - y)^2}}{c} \quad ; \quad n = 1,2,\ldots, N \tag{3.1}$$

where N is the number of microphones. The microphone with largest τ_n is taken as reference τ_{ref} and DSB output is found out using Eq. 2.4. and Eq. 2.3. The SRP is given by:

$$\begin{aligned} &V(I) = \int_{-\infty}^{\infty} Y_I(\omega)\, Y_I(\omega)^* d\omega \\ &i.e; \\ &V(I) = \int_{-\infty}^{\infty} \left(\sum_{n=1}^{N} \sum_{q=1}^{N} X_{n,I}(\omega)\, X_{q,I}(\omega)^* e^{-j\omega(\tau_n - \tau_q)} \right) d\omega \end{aligned} \tag{3.2}$$

$V(I)$ is the the SRP at I. Y_I is the Fourier Transform of DSB output while it beamforms at the location I. The source is then detected by thresholding of $V(I)$. If the number of sources is assumed to be K, then the set of points source location can be identified by[25]:

$$\vec{P}_k = argmax\{V(I), k\}; \qquad k = 1,2,\ldots, K \tag{3.3}$$

where $argmax\{\cdot, k\}$ gives I for the k^{th} maximum value.

3.2.1. SRP – PHAT β

During the propagation of sound higher frequencies are more attenuated than lower frequencies. This is characterized by a tilt in the magnitude response of the room transfer function. This means that SRP computation in Eq. 3.2 is dependent on the spectral coloring of the source and the room impulse response. But for SRP to be a better indication of the location of a source it should be made independent of the spectral magnitude and more sensitive to the phase. This can be done using a PHAT whitening filter[6][12]. The PHAT filter is given by Eq. 5.28

But the conventional PHAT transform also tends to amplify the noise level if the SNR is less than 0dB over a large spectral region. This problem can be addressed by performing

controlled/partial whitening. Parameterized PHAT, referred to as PHATβ [6][7] can be used for this.

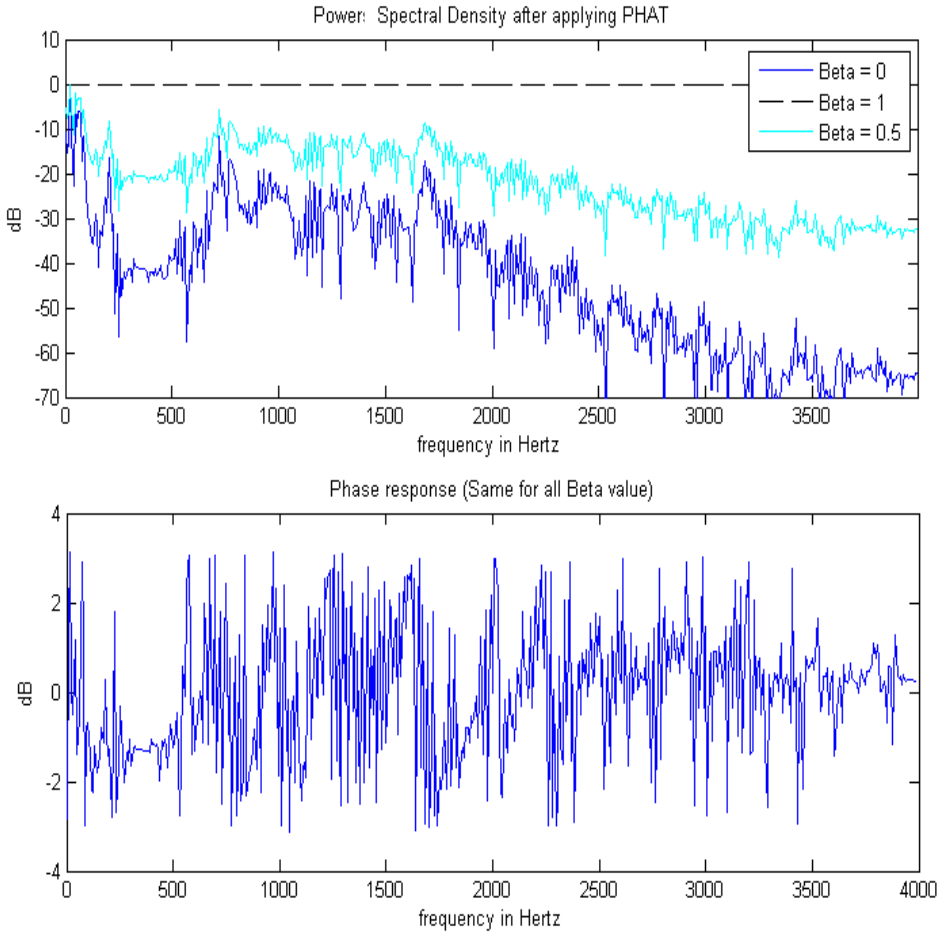

Figure 3.1: PSD and Phase Response after PHATβ

The effect of β on the signal spectrum is shown. β = 1 whitens the spectrum completely. Partial whitening is obtained when β = 0.5. The Phase response is preserved in all instances.

PHATβ is given by[6] :

$$\Theta_{n,\beta}(\omega, I) = \frac{X_n(\omega, I)}{\left|X_n(\omega, I)\right|^{\beta}} \quad ; \quad 0 \le \beta \le 1 \tag{3.4}$$

where $X_n(\omega)$ is the Fourier Transform of $x_n(t)$; the signal received at each of the array elements. β is the whitening parameter. Conventional PHAT is obtained for β = 1. β= *0*

means no PHAT is performed. β can be varied in the range 0 to 1 to obtain various levels of whitening. Figure 3.1 shows the effect of PHAT- β on the spectrum. Substituting from 3.4 in 3.2 and including constant weights (A_n , A_q) , SRP – PHATβ is given by :

$$\check{V}_{\beta}(I)=\int_{-\infty}^{\infty}\left(\sum_{n=1}^{N}\sum_{q=1}^{N}A_n A_q \Theta_{n,\beta}(\omega, I)\,\Theta_{q,\beta}(\omega, I)^{*}\right)d\omega \qquad (3.5)$$

The signal at the closer microphones are weighted more than farther ones. Inverse of the distance from the target point is used as the weight. They are normalized such that the closest microphone will have a weight of one.

$$A_n=\frac{\min\left(\|\vec{r}_s-\vec{r}_n\|\right)}{\|\vec{r}_s-\vec{r}_n\|} \qquad (3.6)$$

3.2.2. SRCP – PHAT β and CFAR Thresholding

$\check{V}_{\beta}(I)$ is a representation of the AS as it gives an indication of possible locations of the sound source. Higher $\check{V}_{\beta}(I)$ indicates the presence of a sound source. A threshold must be applied to the SRP image to detect the presence of a sound source at a given grid point *I*. A Constant False Alarm(CFAR) threshold based on negative peaks of Steered Response Coherent Power (SRCP)[13] is used.

SRCP is a slight modification to SRP – PHAT β and is given by [13]:

$$\tilde{V}_{\beta}(I)=\int_{-\infty}^{\infty}\left(\sum_{n=1}^{N}\sum_{q\neq n}^{N}A_n A_q \Theta_{n,\beta,I}(\omega)\,\Theta_{q,\beta,I}(\omega)^{*}\right)d\omega; \qquad (3.7)$$

Figure 3.2 illustrates an SRCP image after partial whitening (β = 0.7). Field of View is from 0-3.6m in both x and y direction. A spatial resolution of 0.04m is used resulting in a 91 x 91 array of grid points. In computing SRCP the autocorrelation terms are subtracted out. This makes it possible for SRCP to be negative also. Negative areas in the SRCP image indicate an incoherent summation and represent noise. These points are used to statistically model the noise[6] .

The positive peaks $\tilde{V}_{\beta}(I_p)$ are possible candidates to represent a source. $\tilde{V}_{\beta}(I_{pj})$, the negative values in the neighborhood of $\tilde{V}_{\beta}(I_p)$ are used for determining the threshold.

The neighborhood is defined by the grid points in a square(2D) or a cube (3D) with I_P as the center.

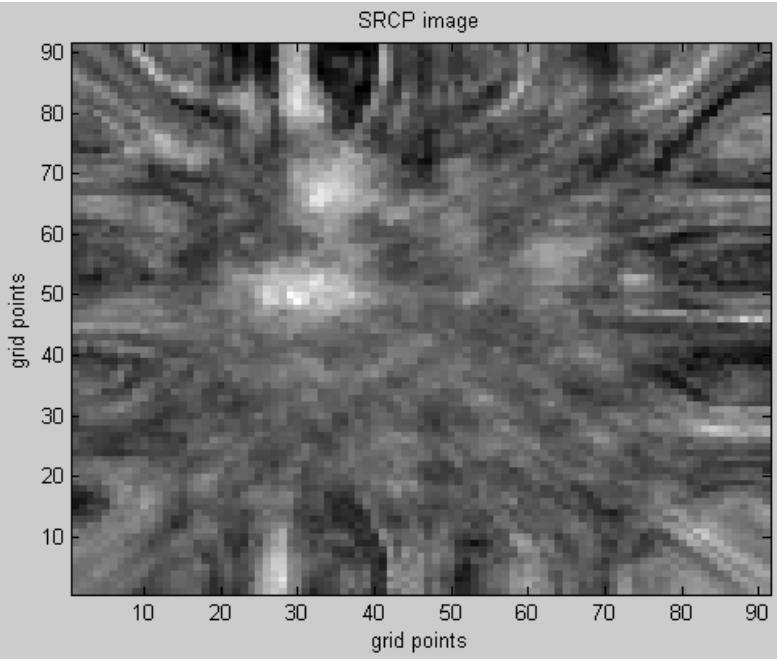

Figure 3.2: SRCP -PHAT β

SRCP image of the FOV with β = 0.7 intensity is represented as a scale from black to white.

Noise is modeled using the Weibull distribution[14]. The probability of a False Alarm(FA) is given by :

$$P_{FA} = 1 - \exp\left(\left(\frac{T_{FA}}{a}\right)^{b}\right) \tag{3.8}$$

where T_{FA} is the threshold for given rate of FA, a is the scale parameter and b is the shape parameter. The value of b is dependent on the actual geometry of the array. a is estimated from the local statistics as:

$$\hat{a} = \left(\frac{1}{N}\sum_{j} \tilde{V}_{\beta}(I_{pj})^{b}\right)^{\frac{1}{b}} \tag{3.9}$$

The value of a estimated is used to find T_{FA} from Eq. 3.8 :

$$T_{FA} = -\hat{a} \ln\left(\frac{1}{P_{FA}}\right)^{\frac{1}{b}} \tag{3.10}$$

Now the sound source is detected using a soft- thresholding.

$$\gamma(I_p) = \begin{cases} \tilde{V}_\beta(I_p) - T_{FA} & ; V_\beta(I_p) \geq T_{FA} \\ 0 & ; V_\beta(I_p) < T_{FA} \end{cases} \tag{3.11}$$

γ(I) acts as the detection statistic for the source. Higher values indicate greater probability of finding a source. Let **P** represent the set of all possible candidates where there exists a source.

$$\begin{aligned} \boldsymbol{P}_w &= \left\{ \vec{\phi}(I_p, t_u) \ : \ \gamma(I_p) > 0 \ \text{ and } \ u = w \right\} \\ \boldsymbol{P} &= \bigcup\nolimits_w (\boldsymbol{P}_w) \end{aligned} \tag{3.12}$$

where $\vec{\phi}$ is a vector with space and time dimensions. $\tilde{V}_\beta(I)$ is computed in overlapping time windows. This results in a sequence of ASes indexed by w and center time denoted by t_w. The set of detections in w^{th} AS is denoted by $\boldsymbol{P}_w$. Figure 3.3 shows the sources estimated from the SRCP image shown in Figure 3.2. A P_{FA} of 6.04 x 10^{-5} (1/(91 x 91)) corresponding to one in every two frames is used. b = 1.26.

The thresholding performed is the lowest level of scene segmentation where the pixels in AS which do not represent a source are rejected and a set of all possible sound sources is defined. The following stages of Scene Segmentation refine the set $\boldsymbol{P}_w$ and tag the remaining elements with a stream ID.

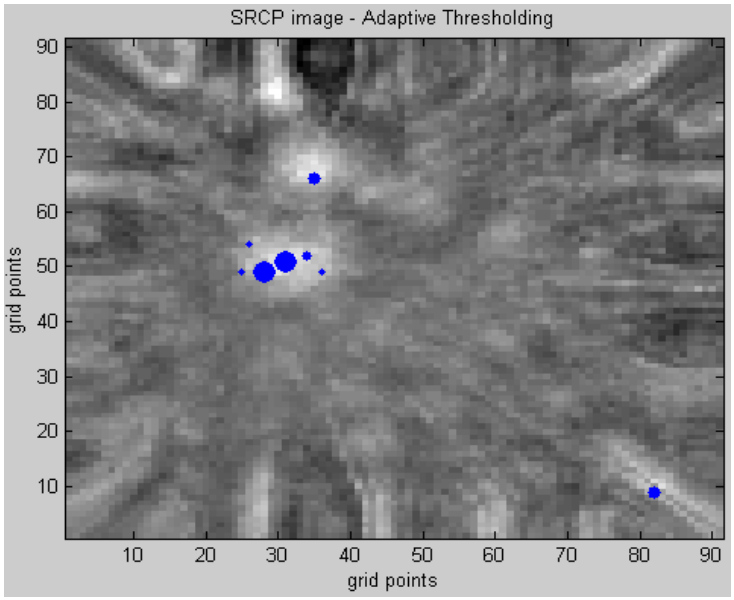

Figure 3.3: SRCP image with adaptive thresholding

The solid circles represent detections. The radius of the spot is scaled according to the confidence of detection.

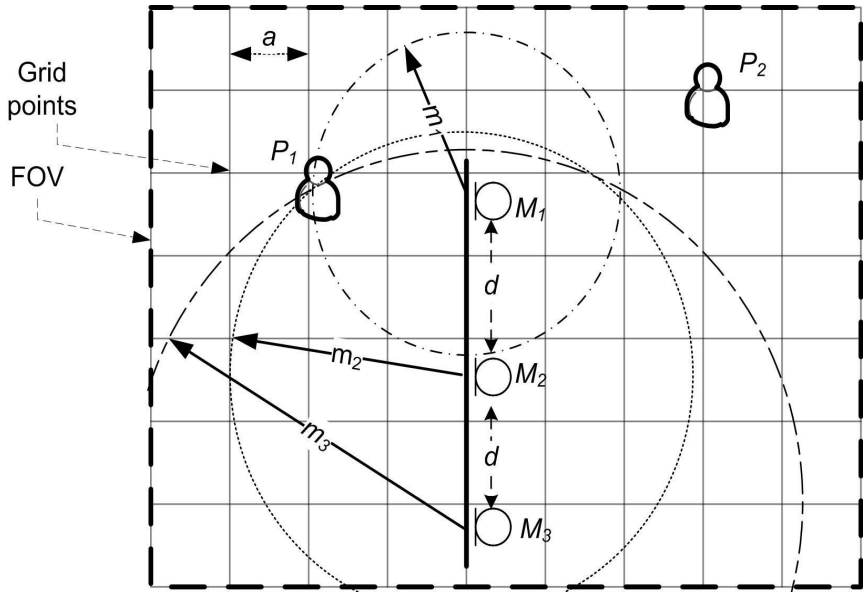

Figure 3.4: Setup for Sound Source Localization using SRP

Each intersection corresponds to a grid point. P_1 and P_2 are sound sources. d is the inter microphone spacing. a is the inter grid spacing. m_1, m_2 and m_3 are the distances from P_1 to M_1 M_2 and M_3.

3.2.3. Design Issues

Figure 3.4 shows the basic setup for SSL using SRCP–PHATβ . P_1 and P_2 are the sources present. The intersection of horizontal and vertical lines represent a grid point. d is the inter microphone spacing and a is the inter grid spacing. The factors to be decided are a, d, β and the number of microphones in the array N.

Inter- grid distance a :

The grids must be close enough so that irrespective of the target's actual position, it will be approximated to the nearest grid point. The inter grid spacing is given by[6] :

$$Q(a) = sinc\left(\frac{2\pi (f_h / f_s)\sqrt{D}\, a}{2}\right) \tag{3.13}$$

where $Q(a)$ is the power loss due to grid quantization. f_h is the highest frequency in the target signal and f_s is the sampling frequency for discrete processing. D is the number of dimensions.

Inter- microphone spacing d :

d is a design parameter of the DS beamformer and section 2.3 explains the effect of d on the performance of DSB. $d < c / f_h$ may be considered as the design constraint to avoid spatial aliasing. But spatial aliasing will not be occurring irrespective of d as [15] :

1. The source in the FOV is present at the near field rather than the far field as assumed in section 2.3.
2. The higher frequencies present in the speech signals enhances its harmonically related lower frequencies. Hence even the frequencies above the cut off would enhance the directionality.

Higher d is desirable as it increases the array aperture and ensures uniform coverage of FOV which is located in the near field. Hence the size of FOV and physical realizability are the factors governing d.

Number of microphones N:

The minimum number of microphones required for SSL is 3 for a 2D FOV and 4 for 3D FOV. For DSB there is a 3dB increase with every doubling of number of microphones[16]. In the experiments for this thesis 16 microphones are used.

Whitening parameter β :

β can take a value from 0 to 1. The optimum value of β has been suggested after simulation studies in [6] and experimental studies in [12]. For human speakers in an office room environment β ranging from 0.65 – 0.7 is found to be giving optimum performance in sound source detection.

3.3. **Conclusion**

There are broadly TDOA based and SRP based SSL techniques. The SRP technique outperforms TDOA when multiple sources are involved. Whitening tends to improve SRP performance. SRP – PHATβ[6] a partial whitening method was reviewed. An adaptive thresholding of the SRCP(modification of SRP) image is done to obtain the source locations.

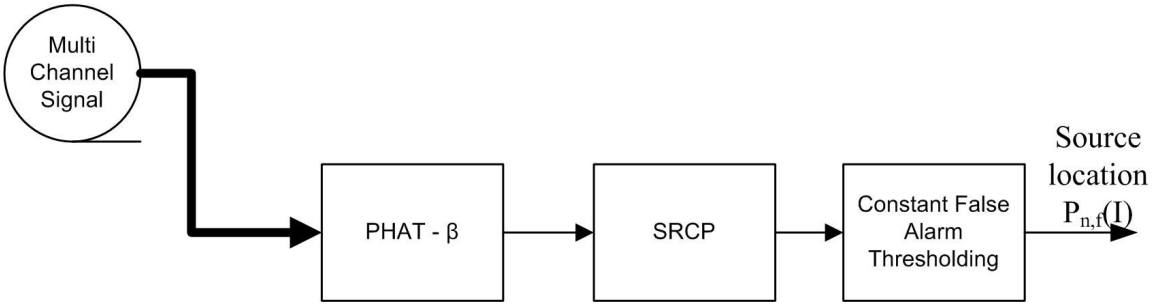

Figure 3.5: SSL using SRP PHAT β and CFAR

Figure 3.5 shows an overview of SSL using SRP PHAT β . Individual channels are pre-filtered to obtain the desired amount of spectral whitening. Then SRCP is found using a DS steering array. A Constant False Alarm(CFAR) thresholding is done to detect sound sources and estimate their location. It should be noted that the goal of this system is not

the enhancement of sound. The source location obtained here can be used for spatial filtering and enhancement of the source signal.

SSL results in multiple detections across ASs. There can be false detections as well as multiple detections of the same source. These must be removed to the maximum possible extent and the remaining detections be linked together to result in streams.

Chapter 4. Audio Scene Segmentation Using Spatial Cues

After thresholding and estimating the sources present in a time frame the next task in ASA is to link the sound sources across time. This chapter presents an attempt to link sources using proximity in spatial location. This is the simplest approach to ASS where the sources are tracked across time. After that a beamformer can be set up to focus on one speaker at a time to obtain the stream. Figure 4.2 depicts a part of a scene. Contiguous AS linked together by streams form a scene. See section 1.1. for more rigorous definition of the terms.

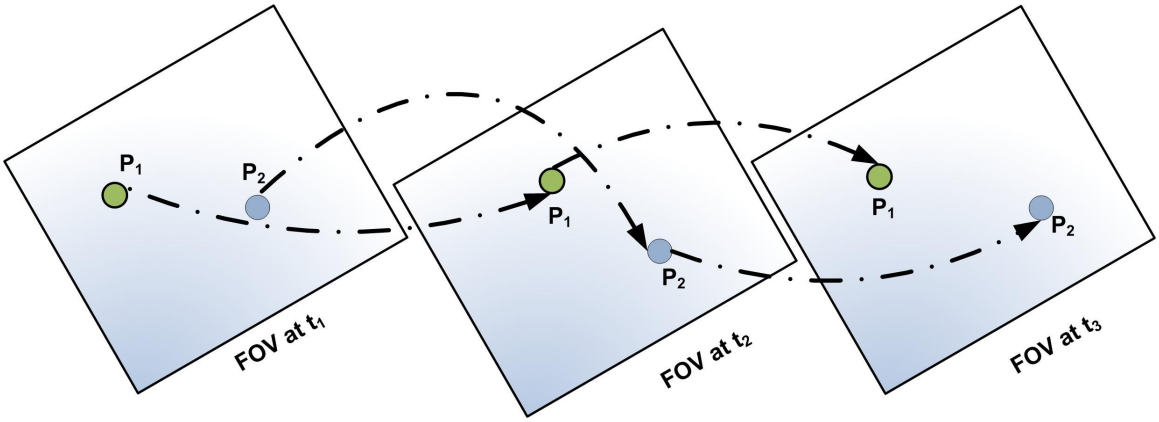

Figure 4.1: Concept of Stream and Audio Scene Segmentation.
P_n represent the sources and the black line represents a stream.
Each rectangle is an AS of FOV at time window t_n.

4.1. **Mathematical Model**

First a metric to measure proximity is defined. Spatial proximity is measured using norm-2 distance. The norm-2 spatial distance between any two detections in space $\vec{\phi}(I,t)$, is given by :

$$\left\|\vec{\phi}(I_i,t_u)-\vec{\phi}(I_j,t_v)\right\|_S \stackrel{\text{def}}{=} \|I_i-I_j\| \stackrel{\text{def}}{=} \sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \tag{4.1}$$

where $\boldsymbol{P_u} \ni \vec{\phi}(I_i,t_u), \boldsymbol{P_v} \ni \vec{\phi}(I_j,t_v)$

The temporal proximity is measured using:

$$\left\|\vec{\phi}(I_i, t_u) - \vec{\phi}(I_j, t_v)\right\|_T \stackrel{\text{def}}{=} \left|t_u - t_v\right| \tag{4.2}$$

where i, j denotes the sources and u, v denotes ASes. $\vec{\phi}(I_i, t_u)$ and $\vec{\phi}(I_j, t_v)$ are considered as belonging to the same stream if they are in space-time proximity. They belong to same stream for :

$$\begin{aligned} &\left\|\vec{\phi}(I_i, t_u) - \vec{\phi}(I_j, t_v)\right\|_S < \rho \quad \text{and} \\ &0 < \left\|\vec{\phi}(I_i, t_u) - \vec{\phi}(I_j, t_v)\right\|_T < \psi \end{aligned} \tag{4.3}$$

where ρ , ψ are the spatial and temporal thresholds.

4.2. **Removal of Secondary detections**

Sound Source Detection (SSD) sometimes results in multiple detections of the same source. Before the sources are linked across time these must be removed. First the detection with highest $\gamma(I_p)$ is found. Then any detection within the distance of ρ_0 from it is taken as a secondary detection of the same source. Hence they are dropped. Then the detection with next higher $\gamma(I_p)$ is searched for and any of its secondary detections are dropped. This process is continued until all the detections in the AS are verified. The same process is done for all AS. This results in set $\boldsymbol{G}$.

$$\begin{aligned} &\boldsymbol{G_w} = \left\{\vec{\phi}(I_i, t_w) \;:\; \|\vec{\phi}(I_i, t_w) - \vec{\phi}(I_j, t_w)\|_S > \rho_0 \quad \forall\, i \neq j\right\} \\ &\text{also } \boldsymbol{G_w} \subset \boldsymbol{P_w}; \boldsymbol{G} = \bigcup_w \left(\boldsymbol{G_w}\right) \end{aligned} \tag{4.4}$$

Figure 4.2 shows the flow chart for the removal of secondary detections. Each element of $\boldsymbol{G}$ is represented by $G[w][i]$ where the first index correspond to the AS and the second index correspond to the detection.

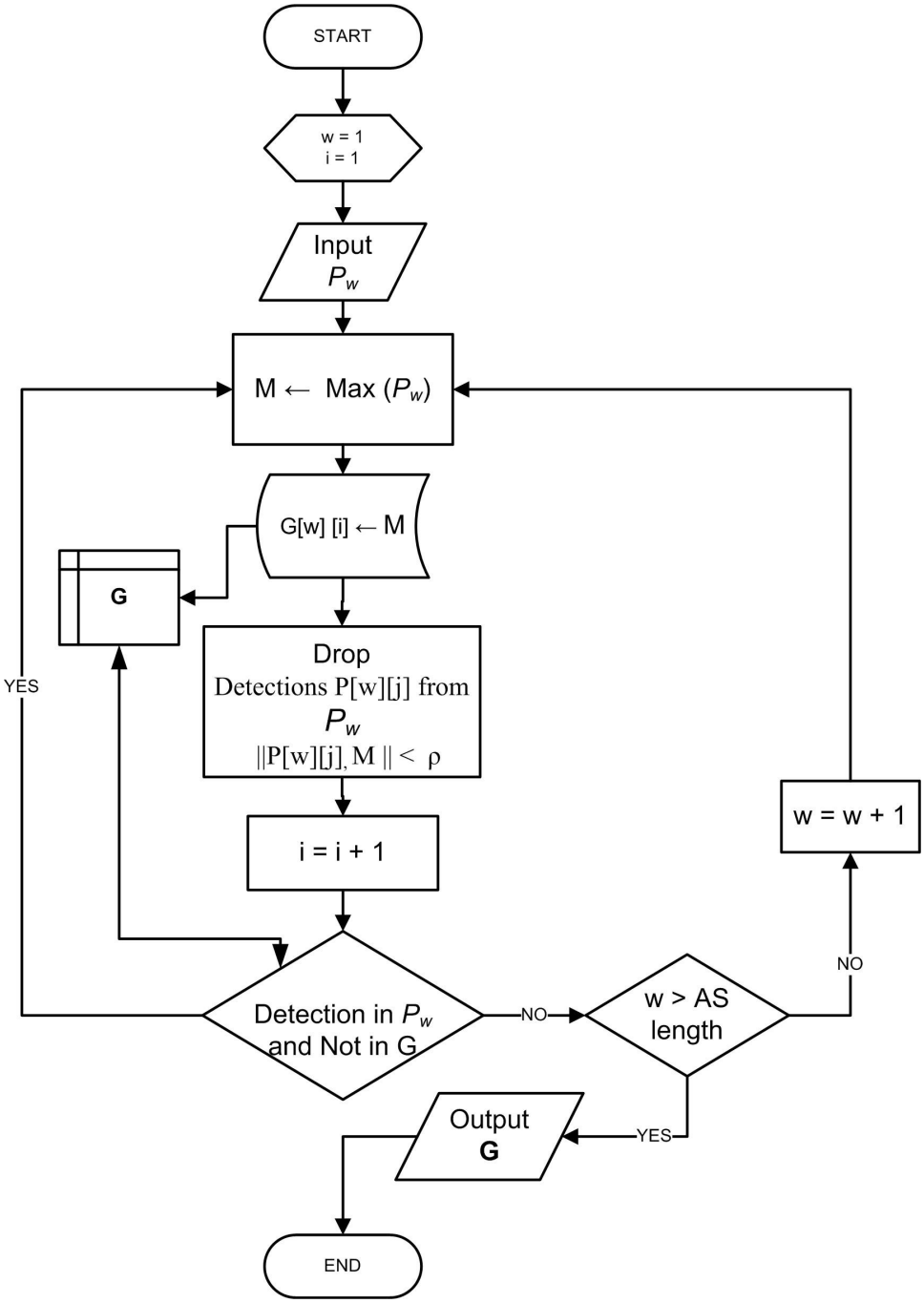

Figure 4.2: Removal of secondary detection of the same source in one AS.

4.3. **Linking Detections across AS**

Here each element in $\boldsymbol{G}$ is assigned a stream ID such that the elements which have the same stream ID are said to be the member of same stream. The detections are processed sequentially. $\boldsymbol{G}$ obtained in the previous process is read in. w,i are the indexes for AS and the detections. All detections in $\boldsymbol{G}$ which are within temporal distance of ψ of the current point $\vec{\phi}(I_i, w)$ are checked for spatial proximity so that :

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{w},\boldsymbol{i}} &= \left\{ \vec{\phi}(I_j, t_w + t_m) \ : \ \|\vec{\phi}(I_j, t_w + t_m) - \vec{\phi}(I_i, t_w)\| < \rho \right\}; \\ \boldsymbol{G} &\ni \vec{\phi}(I_i, t_w), \boldsymbol{G} \ni \vec{\phi}(I_j, t_w + t_m) \end{aligned} \tag{4.5}$$

$m = 1,2,\ldots,\psi - 1$ and $j = 1,2,\ldots, N_{w+m}$. N_w is the number of detections in the w^{th} AS.

$\boldsymbol{C}_{\boldsymbol{w},\boldsymbol{i}}$ is a set of detections which are linked to $\vec{\phi}(I_i, t_w)$. It must be ensured that no detection is linked to two previous points. i.e

$$\boldsymbol{E} = \boldsymbol{C}_{\boldsymbol{n},\boldsymbol{l}} \cap \boldsymbol{C}_{\boldsymbol{p},\boldsymbol{m}} = \emptyset \quad \forall\, l \neq m \ \& \ n \neq p \tag{4.6}$$

This is ensured by using the minimum distance measure.

$$(f,k) = argmax_{n \backslash p, l \backslash m} \left\{ \left\| \vec{\phi}(I_j, t_f) - \vec{\phi}(I_l, t_n) \right\|_S, \left\| \vec{\phi}(I_j, t_f) - \vec{\phi}(I_p, t_m) \right\|_S \right\} \tag{4.7}$$

where $\vec{\phi}(I_j, t_f) \in \boldsymbol{E}$

$$\boldsymbol{C}_{\boldsymbol{f},\boldsymbol{k}} = \boldsymbol{C}_{\boldsymbol{f},\boldsymbol{k}} - \{\vec{\phi}(I_k, t_f)\} \tag{4.8}$$

Figure 4.3 illustrates the flowchart for linking detections across ASS. The stream ID of each element in $\boldsymbol{G}$ is stored at *StrID*[*w*][*i*]. *Dis*[*w*][*i*] has the distance from the previous detection in the same stream. If the current point $\vec{\phi}(I_i, t_w)$ (*G*[*w*][*i*]) is the origin of a stream, *Dis*[*w*][*i*] is set to infinity(very large number). A value of 'Null' for *StrID*[*w*][*i*] means that no stream ID has been assigned for *G*[*w*][*i*] yet. This results in streams defined by :

$$\boldsymbol{H}_{\boldsymbol{\zeta}} \stackrel{\text{def}}{=} \left\{ \vec{\phi}(I_i, t_w) : \begin{matrix} \|\vec{\phi}(I_i, t_w) - \vec{\phi}(I_j, t_f)\|_S < \rho \\ 0 < \|\vec{\phi}(I_i, t_w) - \vec{\phi}(I_j, t_f)\|_T < \psi \end{matrix} \ ; \ \begin{matrix} \vec{\phi}(I_i, t_w) \in \boldsymbol{G} \\ \vec{\phi}(I_j, t_f) \in \boldsymbol{G} \end{matrix} \right\} \tag{4.9}$$

where $\zeta = 1,2,3,\ldots, N_{str}$; N_{str} is the number of streams.

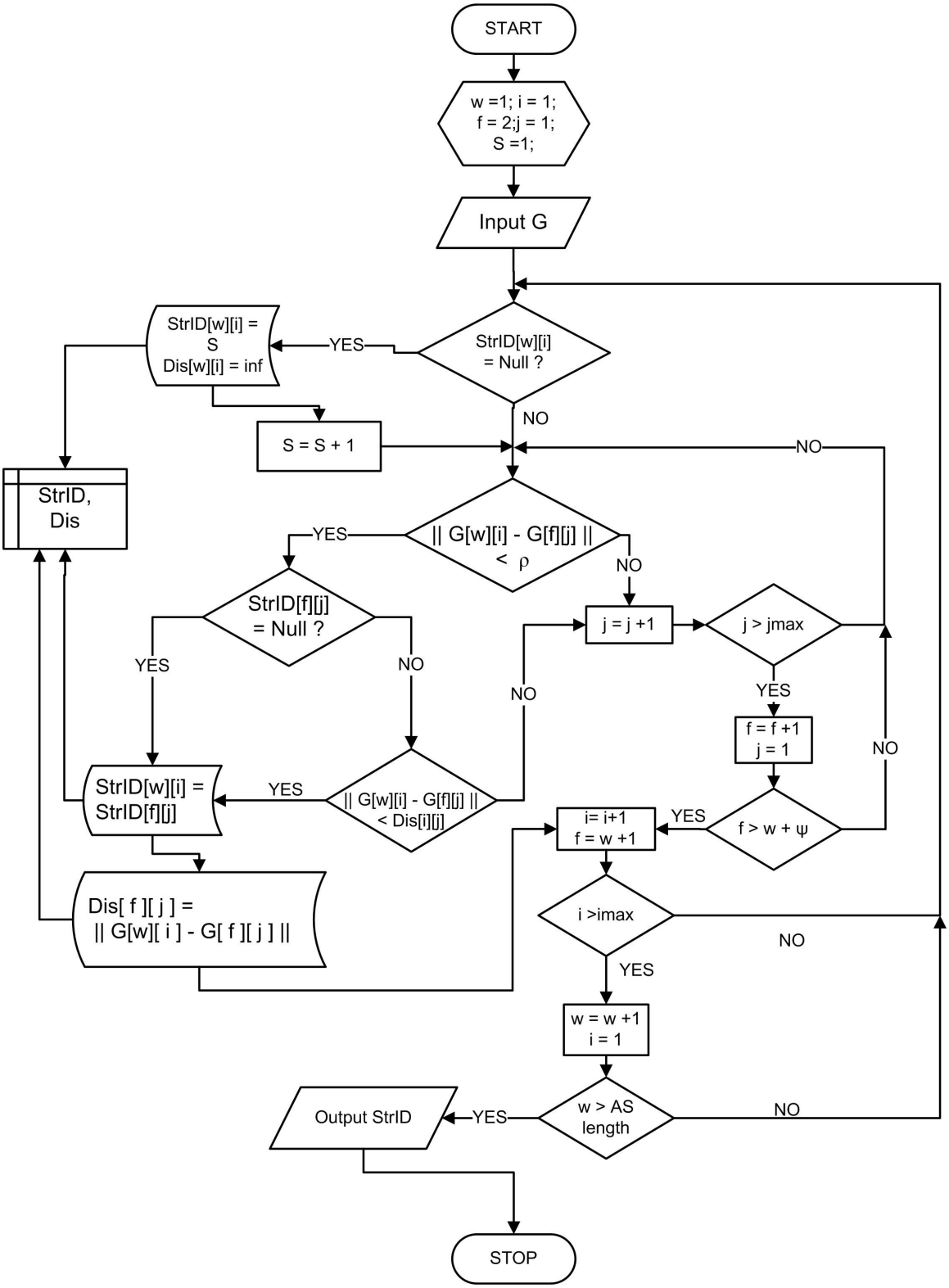

Figure 4.3: Flowchart for ASS using spatial cues

4.4. **Experiment Setup**

The experiment was set up in a typical office space. An array of 16 microphones was used. Two male speakers were made to read out different printed texts while moving in a predefined hexagonal path. The microphones were placed in the perimeter of a 3.6 m. X 3.6 m. square which circumscribes the speakers' paths. Microphones were at a height of 1.5 m. from the floor. Figure 4.4 shows the experimental setup. The gray bars represent the acoustic foam panels used to reduce the reverberations from the walls. The colored circles define the path for each speaker. The dark color represents speaker1 and light color represents speaker2. At each spot, the speakers are made to read out simultaneously for 3 seconds. Then they move to the next spot within the next 3 seconds. The speakers' position for time 3-6, 9-12, 15-18, ... seconds are hence known. Initial 3 seconds are used for speakers to settle down. Speaker1 is made to move in clockwise direction while speaker2 in anti-clockwise direction.

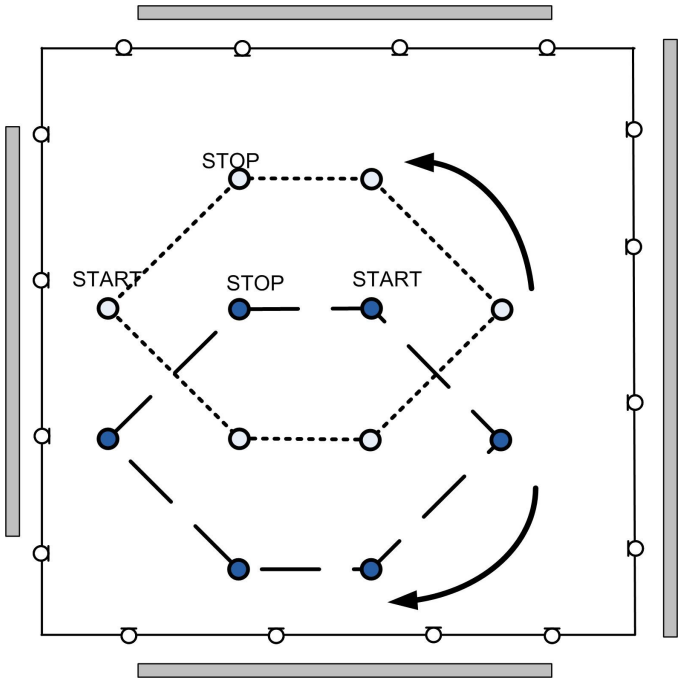

Figure 4.4: Experiment setup for ASS using spatial cues.

Table 4.1. lists the location of speakers at various times after the recording has started. '-' denotes that the speaker location is unknown. The sound sources were approximately at a height of 1.5m.

Table 4.1: Predefined speaker locations.(in meters;ordered pair(x,y))

Time (sec.)	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24	24-27	27-30	30-33	33-36
Speaker 1	-	2.0,2.0	-	2.8,1.2	-	2.0,0.4	-	1.2,0.4	-	0.4,1.2	-	1.2,2.0
Speaker 2	-	0.4,2.0	-	1.2,1.2	-	2.0,1.2	-	2.8,2.0	-	2.0,2.8	-	1.2,2.8

The scene was recorded using 16 microphones at a sampling frequency of 22.05 kHz and digitally stored for further processing. The noise sources include air conditioner vents, CPU fans and sound of traffic through the windows. Also while locating one speaker the other speaker acts as noise. The recordings are done with the help of a Delta 1010™ sound-card. The microphones are phantom powered by Audio Buddy™ pre-amplifiers. The apparatus details are listed in Table 4.2.

Table 4.2: Apparatus details for experimental evaluation of ASS using spatial cues

Equipment	Details
Microphone	Behringer© ECM8000 [17], condenser type, Omni directional, Frequency response:15Hz to 20 kHz .
Acoustic Foam Panels	Auralex MAX-Wall™ [18], Noise Reduction Coefficient - 1.05
A/D converter	M-Audio Delta1010™ [19] Digital recording system (2 Nos.), Frequency response:20Hz – 22kHz, 8 X 8 analog I/O
Pre-amplifier	M-Audio Audio Buddy™ [20], 2-channel, Phantom power, Frequency response :5Hz – 50kHz
Software	Jack audio connection kit 0.3.2 [21], Ubuntu studio 8.04

The 16 channel audio data is recorded for 36 seconds. The data is then processed off-line. The processing involved are pre-whitening(Eq.3.4), finding SRCP(Eq.3.7) CFAR

thresholding (Eq.3.11 and Eq.3.12), and ASS(Figure 4.2 and Figure 4.3). The processing parameters are listed in Table.4.3

Table 4.3: Processing parameter for experiment (ASS using spatial cues)

Parameter	Value
Whitening Parameter - β	0.7
Inter Grid Spacing - a	0.04m
Processing Window	4.0×10^{-3}s. With 50% overlap. (50 AS / sec)
Microphone Geometry	Perimeter with inter-microphone spacing of 0.81 m
Bound for the Neighborhood - r	7 grid points , 0.28m
Probability of False alarm P_{FA}	1 False alarm per 2 AS; 1/(Number of grid points) = $1/(91 \times 91 \times 2) = 6.04 \times 10^{-5}$
Shape Parameter - b	1.26
Minimum length for a valid stream	20 AS (0.4 seconds)

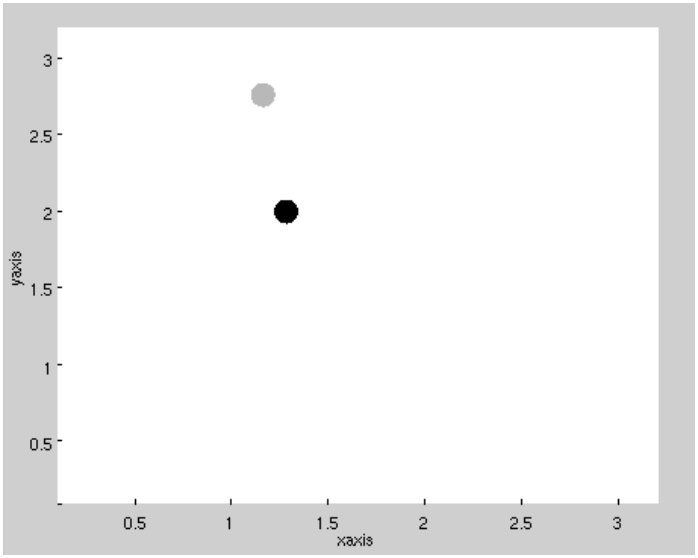

Figure 4.5: AS representation after scene segmentation.
The solid circles with different shades represent different speakers. Here two speakers were detected.

Figure 4.5 shows an AS after ASS is carried out. It shows two detections roughly at (1.2,2.0)m and (1.1,2.8)m. This AS is taken at a time instant of 34 seconds. It corresponds approximately with predefined location given in Table 4.1

4.5. **Performance Analysis**

ASS was carried out for varying values ρ and ψ . The performance metric is defined as :

$$\chi(\rho,\psi) \stackrel{\text{def}}{=} \left\{ \begin{array}{ll} \dfrac{\tilde{N}_{TD}/N_{TD}}{\tilde{N}_s/N_s} & ;\ \tilde{N}_s \geq N_s \\ 0 & ;\ \tilde{N}_s < N_s \end{array} \right\} \qquad (4.10)$$

where $\tilde{N}_{TD}$ is the number of true detections obtained experimentally. N_{TD} is the maximum number of true detections achievable. $\tilde{N}_S$ is the number of streams resulted because of segmentation and N_S is the number of streams actually present. In the experiment the segmentation should ideally result in two streams (one for each speaker). i.e. $N_S = 2$. The maximum number of true detections achievable is equal to the true detection achieved after SRCP – PHATβ and removal of secondary detections. i.e. N_{TD} is the number of elements in $\boldsymbol{G}$ (Eq. 4.4). All the values are estimated only during the duration where the speaker locations are known (Table 4.1). Figure 4.6 shows χ as a function of ρ and ψ . It can be seen that the performance is not tightly dependent on the spatial threshold. This is because the speakers were stationary for short intervals of time(< 3 seconds). In this experiment the optimum performance was achieved at $\rho = 7.5\,\text{cm}$. and $\psi = 6\,\text{s}$ At these values there were 28 segments. The number of segments were counted after short segments(< 0.4s.) are dropped.

When speakers change their position while remaining silent, the algorithm is unable to track the source. It is also observed that at every brief period of silence or miss-detection a new stream is created. An over segmentation is results as the algorithm is very sensitive to these factors.

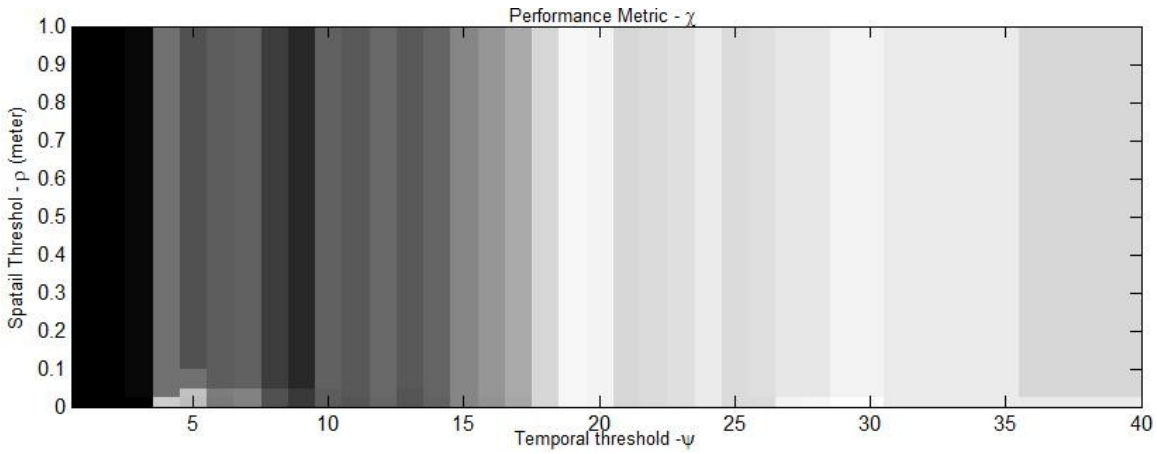

Figure 4.6: Performance of ASS using spatial cues.

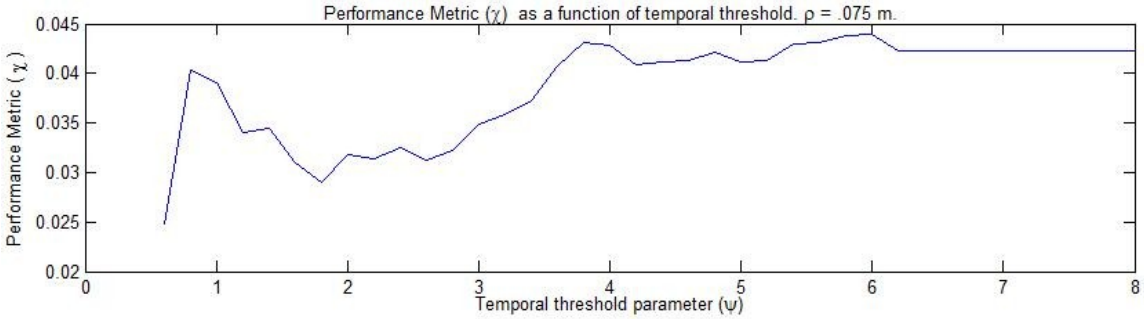

Figure 4.7:Performance of ASS using spatial cues as a function of Ψ at ρ= 7.5 cm

4.6. **Result**

Table 4.4 shows the streams obtained after ASS with ρ= 0.30m. , ψ = 6s. There were 30 streams detected out of which two were false detections. The last column shows the speaker to which the stream actually belonged. This is inferred using the predetermined locations listed in Table 4.1. The difference in the detected and predetermined source locations are due to measurement error in setting up the microphones, marking of the coordinates and grid resolution. It is observed that detections which are within the thresholds also end up in different streams. This happens as false initiation of segment results in parallel streams within the threshold. In that case the detection is classified into the closer stream.

Table 4.4: Streams Detected at ρ= 0.30m. , ψ = 6sec.

Time ID	Time	Stream ID	Median. Spatial Coordinate	Speaker1/2
1	3 – 6	1	2.04, 2.04	Speaker 1
		2	2.00, 2.04	Speaker 1
		3	2.00, 2.08	Speaker 1
		4	0.40, 2.00	Speaker 2
		5	0.40, 1.96	Speaker 2
		6	0.44, 1.96	Speaker 2
		7	1.60, 1.60	*False Detection*
2	9 – 12	8	1.24, 1.32	Speaker 2
		9	2.68, 1.20	Speaker 1
		10	2.72, 1.20	Speaker 1
		11	2.74, 1.16	Speaker 1
		12	2.88,2.92	*False Detection*
		13	2.72, 1.16	Speaker 1
3	15 – 18	14	1.96, 1.36	Speaker 2
		15	1.92,0.52	Speaker 1
		16	1.88,0.52	Speaker 1
4	21 – 24	18	2.64, 2.08	Speaker 2
		19	1.16, 0.56	Speaker 1
		20	1.12, 0.56	Speaker 1
5	27 – 30	21	1.96, 2.72	Speaker 2
		22	0.60, 1.24	Speaker 1
		23	1.92, 2.72	Speaker 2
		24	1.96, 2.76	Speaker 2
		25	0.60, 1.28	Speaker 1
6	33 – 36	26	1.12, 2.80	Speaker 2
		27	1.20, 1.96	Speaker 1
		28	1.16, 2.76	Speaker 2
		29	1.24, 1.96	Speaker 1
		30	1.24, 1.96	Speaker 1

4.7. **Conclusion**

In a 36 second recording of two speakers with about 18 seconds of active speech in it, 28 streams is a case of over segmentation. The number of streams would have been higher if shorter segments (< 20 AS) were not dropped. This demonstrates a need for finding other robust features for performing ASS. The features to be used must be speech-invariant and must be dependent on the speaker. Some features are analyzed for these characteristics in

the coming chapter. These features can be used to do a second pass combining the localized streams. The processing thus far cannot be considered as streaming. For the streaming process to be complete the sources with the same stream ID must be enhanced using beamforming and then linked together.

Chapter 5. Auditory Features for ASS

5.1. Introduction

Chapter 4 demonstrated the need of using auditory features for performing ASS. Spatial cues alone could not give an acceptable level of performance. Since the speaker locations are known, they can now be beamformed on and their auditory features extracted. The task of grouping the localized streams essentially becomes a speaker recognition task. The problem is easier than standard speaker recognition as the number of candidates will be limited (2 in this thesis). But the streaming system is not trained on any particular speaker and therefore does not have *a priori* statistical knowledge about speaker features. Also the beamformed signal will have interference from other speakers. As the recordings are from distant microphones, the resulting feature will be degraded by the room modes and reverberations.

In this chapter possible features and their combinations are analyzed using single microphone clean speech recordings. They are tested to assess recognition performance on text independent speech. The feature or combination of features which give better recognition rate will then be used on the beamformed signals to perform ASS in the coming chapters.

5.2. Audio Features ; Mathematical Models

5.2.1. Pitch

Pitch is defined as the perceived fundamental frequency of a sound[22]. The auditory system can perceive the pitch of a complex tone even when the fundamental frequency is actually missing [23]. The algorithms to estimate the pitch can be broadly classified as place(spectral), time, and place-time approaches[2]. Spectral methods include harmonic sieves[24] and partial frequency histograms. Time domain methods extract the periodicity information from the autocorrelation of the signal. In the place-time approach the signal- is passed through a filter bank and the outputs are analyzed temporally and spectrally. In

[25] place-time approach is used for multiple pitch and vowel estimation for simultaneous utterance.

The spectral approach suffers from its dependence on analysis window shape and duration. The place-time method gains over the temporal method as it allows to undo any amplitude mismatches between spectral regions before detecting periodicity in time[3]. For example, in the spectro-temporal approach the weights of each frequency channel can be adjusted to perform "spectral whitening". But the disadvantage of these methods is that they are computationally expensive as they try and model the hearing system using filter banks. A computationally less expensive way is proposed in [26] where the signal is divided into two channels; one less than 1kHz and the other greater than 1kHz . Taking the collapsed average of the generalized spectrum after pre whitening[27] also achieves the same goal with lesser computation. [27] and [26] use conventional signal processing tools whereas [25] uses CASA.

In this thesis the pitch estimation as in [27] is used. Consider a sound segment $s(t)$, which is 50ms in duration and sampled to obtain s[n] where $t = n\Delta t$ and Δt is the sampling interval. S[m] represents the FFT of s[n]. Then by [27] the generalized spectrum is defined as:

$$\boldsymbol{G}[m,k] = \mathbf{E}\left\{S[m]\cdot S^{*}[m-k]\right\}$$

$$\text{where } k,\ m = 0,1,2,\ldots,M-1;\ M = \left[\frac{50\text{ms}}{\Delta t}\right] \tag{5.1}$$

$\boldsymbol{G}[m,k]$ is a matrix of order M by M where each row indexed by m and each column indexed by k. The average of $\boldsymbol{G}$[m,k] over m will reveal the periodicity in spectrum. The normalized collapsed average of $\boldsymbol{G}[m,k]$ is obtained by [27] :

$$C[k] = \frac{\sum_{m=0}^{M-1} S[m]S^{*}[m-k]}{\sum_{m=0}^{M-1} S[m]S^{*}[m]} \tag{5.2}$$

C[k] is normalized such that the zero lag (power) is unity. The peaks in C[k] are directly related to the pitch and the resulting harmonics. Hence IFFT of C[k] is taken and the highest peak in the pitch range 12.5ms (80 Hz) to 33ms (300 Hz) is taken as the pitch.

$$n_p = argmax_n\{c[n]\}$$
$$L_p = \frac{1}{n_p \Delta t} \tag{5.3}$$

L_p is the pitch in Hertz. Each feature will be represented by L followed by a subscript representing the feature.

5.2.2. Envelope Power

The power contained in the envelope of the speech signal gives an indication of the loudness of the speaker. Loudness may be a stable parameter especially within a scene or conversation. The squared envelope is obtained by :

$$s_{env}[n] = |\, s[n] + i\,\mathrm{H}(s[n])\,| \tag{5.4}$$

where $H(.)$ represents the Hilbert transform. $s_{env}[n]$ is then down sampled to 1200Hz after anti aliasing. The envelope power is computed by :

$$L_e = \frac{1}{N}\sum_{n=0}^{N-1} S_{env}^2[n]; \qquad \text{where} \quad N\Delta t = 50\,ms \tag{5.5}$$

5.2.3. Rate of speech

The rate of change of the envelope $S_{\Delta env}[n]$ is measured using the mean of log-difference of the envelope. log-difference is used as it is independent of the magnitude. Its value is dependent on the pace at which the speaker is talking.

$$L_r = \frac{1}{N-1}\left|\log_e\left(1+S_{env}[n]\right) - \log_e\left(1+S_{env}[n-1]\right)\right|; \tag{5.6}$$
where $n = 1,2,\ldots N-1$

5.2.4. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature which is commonly used in Automatic Speaker Recognition[28][29] [30]. It is obtained by mapping the cepstral power to the melody(Mel) scale. The Mel scale is an exponential frequency scale which approximates the human perceptual scaling. Calculating MFCC involves the following steps:

1. The cepstral power $S_{dB}[m]$ is computed by :

$$S_{dB}[m]=20\log_{10}(|S[m]|+1) \tag{5.7}$$

where $S[m]$ is the DFT of $s[n]$.

2. The Mel axis is obtained by N regularly spaced points (f_l) from 0 to 4kHz, which are mapped to a Mel scale by:

$$f_m[k]=\log_{10}\left(\frac{f_l[k]}{700}+1\right)2595 \tag{5.8}$$

where f_l and f_m are in Hertz.The numerical equivalent of Eq. 5.8 which is also widely used is :

$$f_m[k]=\log_e\left(f_l\frac{[k]}{700}+1\right)1127.01048 \tag{5.9}$$

3. Then a Hanning overlapping window is used to average and map the linear scale to the Mel scale. The maxima or the center point of the window coincides with $f_m[k]$ and the window length is $f_m[k+1]-f_m[k-1]$ where k is the index of Mel- frequency.

$$S_{mel}[k]=\frac{\sum_{l=f_m[k-1]}^{f_m[k+1]}\left(S_{dB}[l].h_{W_k}\left[l-f_m[k-1]\right]\right)}{W_k} \tag{5.10}$$

where $W_k=\frac{f_m[k+1]-f_m[k-1]}{\Delta t}$; h_{W_k} is the Hanning window of length W_k

$S_{mel}[k]$ is called the Mel Cepstrum.

4. The Discrete Cosine Transform (DCT) of $S_{mel}[k]$ is taken to obtain the MFCC.

$$L_m[n]=\sqrt{\frac{1}{N}}S_{mel}[1]\cos\left(\pi\frac{n}{2N}\right)+\sqrt{\frac{2}{N}}\sum_{k=2}^{N}S_{mel}[k]\cos\left((2k-1)\pi\frac{n}{2N}\right)$$

where $n=0,1,2,\ldots,N-1$

(5.11)

5.2.5. ΔMFCC

First order differential of MFCC is obtained by :

$$L_{\Delta m}[n]=\frac{\sum_{w=1}^{W}MFCC[n+w]-MFCC[n-w]}{2\sum_{w=1}^{W}w^2} \qquad (5.12)$$

where $W = 2$. The edges are truncated to avoid index overflow.

5.2.6. Vocal tract impulse response

Vocal tract can be coarsely modeled as set of coaxial tubes[22]. Each of the tube will have its resonant frequency and can be modeled as a filter with a pair of complex poles. The IIR filter representing the vocal tract is obtained by cascading these filters. For a model of N/2 tubes, there are N poles and combined filter may be written as [22]:

$$H_v(z)=\frac{1}{1-a_1z^{-1}-a_2z^{-2}-\ldots-a_Nz^{-N}} \qquad (5.13)$$

where N = 2(BW +1) ; BW is the bandwidth of the signal expressed in kHz. Since a sampling frequency of 8kHz is used, the signal is band limited to 4kHz. N =10. Here the gain of the filter is not of concern and is kept at unity. i.e. $H_v(z)$ is evaluated over the unit circle. In order to determine the coefficients a_i consider the time domain response:

$$y[n]=x[n]+\sum_{i=1}^{N}a_iy[n-i] \qquad (5.14)$$

where $x[n]$ is the input and y[n] is the output of the filter. If the filter is used as a predictor for a wide sense stationary signal like voiced speech it becomes :

$$\tilde{y}[n]=\sum_{i=1}^{N} a_i\, y[n-i] \tag{5.15}$$

and a_i is estimated for least mean square error. a_i s are the Linear Predictive Coefficients (LPC). The feature vector is obtained as :

$$L_H[n]=\mathrm{DCT}\left\{20\log_{10}\left|H_v(z)\right|\right\}$$
$$\text{where}\quad z=e^{\left(\frac{j2\pi n}{N}\right)};\quad N=50\text{ms}.\Delta t \tag{5.16}$$

where $\Delta t=1.25\,\mathrm{x}\,10^{-4}$s. DCT is used to make the feature points orthogonal. This allows the optimization of number of dimensions by changing the number of DCT coefficients used in the feature vector. The optimum value for feature length is derived in section 5.4. The DCT is computed as in Eq.5.11.

5.2.7. Center of Mass of Vocal Tract Impulse Response

The center of mass of $|H_v[z]|$ is obtained by:

$$L_c=\frac{\sum_z |H_v[z]|\cdot z}{\sum_z |H_v[z]|}$$
$$\text{where } z=e^{\left(\frac{j2\pi n}{N}\right)} \tag{5.17}$$

5.2.8. Gammatone Frequency Cepstral Coefficients (GFCC)

GFCC [28] is functionally similar to MFCC. It maps the spectral energy to a frequency scale which follows the sensitivity of the ear. The signal is passed through a Gamma-tone filter bank. The filter bank crudely models the cochlear response of the human ear. The center frequencies of the filter bank are placed equally in Equivalent Rectangular Bandwidth(ERB) scale. The mapping between linear and ERB scale is given by[31] :

$$f_{ERB}=24.7\log_{10}\left(4.37\frac{f}{1000}+1\right) \tag{5.18}$$

where f is in Hertz. The Gamma-tone filter impulse response is given by [28]:

$$G_i(t) = at^{(r-1)}\cos(2\pi f_i t + \phi)e^{(-2\pi bt)} \tag{5.19}$$

$$p_n[t] = G_n[n] * s[n] \tag{5.20}$$

where b, ϕ, r, a are bandwidth, phase correction, order and amplitude respectively. f_c is the center frequency of the i^{th} filter. The filter bank outputs N cochlear channels p_i where $i = 1,2,\ldots,N$. Each of the frequency channel P_i is down sampled to 100Hz to obtain a feature vector every 10 ms. P_i is loudness compressed using the cube root function to obtain Gamma-tone Feature(GF). GFCC is obtained by taking the DCT of the resulting signal.

$$G_i[n] = |p_{(downsample)}[n]|^{1/3} \tag{5.21}$$

$$L_G[j] = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} G_i[n]\cos\left(\frac{j\pi}{2\mathrm{N}}(2\mathrm{n}+1)\right); \tag{5.22}$$

where $j = 0,1,\ldots,N-1$

The GFCC is obtained at every 10 ms.

5.2.9. ΔGFCC

The first order differential of GFCC is obtained by :

$$L_{\Delta G}[j] = \frac{\sum_{w=1}^{W} GFCC[j+w] - GFCC[j-w]}{2\sum_{w=1}^{W} w^2} \tag{5.23}$$

where W = 2. The edges are truncated to avoid index overflow.

5.3. **Data set**

The data set for analyzing the features listed in 5.2. consist of clean speech recording of 3 female and 5 male speakers extracted from the repository[32]. Three different recordings are made for each speaker. Each recording is roughly 20 seconds long. So the

total data set consists of 24 (8 x 3) recordings of approximately 20 seconds duration. They are recorded in a close microphone configuration and sampled at 44.1 kHz.

The analysis of the data will only be valid if the data can be characterized as a stationary signal. Voiced sound can be considered to be stationary over a short window(50 ms.). Hence the unvoiced and silent portions of the signal are first removed from the speech

5.3.1. Removal of voiced and silent speech segments

Voiced speech mainly consisting of glottal waves is characterized by concentration of energy in lower bands of the spectrum whereas in unvoiced speech energy is spread out to higher frequencies also. A simple way to test for the existence of higher frequency components is to find out the Zero Crossing Rate (ZCR). ZCR can hence be used to determine whether the speech segment is voiced or unvoiced. The data is analyzed in 25ms. segments. ZCR is given by :

$$z_r[n] = \frac{1}{2}\sum_{n=1}^{N-1} |\mathrm{sgn}(s[n]) - \mathrm{sgn}(s[n-1])|;$$

$$\mathrm{sgn}(x) = \begin{Bmatrix} 1; x>0 \\ -1; x\le 0 \end{Bmatrix} \text{ and } \quad N = \left[\frac{25\mathrm{ms}}{\Delta t}\right] \tag{5.24}$$

Also it is noticed that the energy content in unvoiced segment is less compared to that of voiced segment. Also using this criteria the silence will also be removed. Hence log-energy is used for verifying whether the given speech segment is voiced or not. The log energy is computed by :

$$E_{\log}[n] = \log_{10}\left(\sum_{n=0}^{N-1} s[n]^2\right) \tag{5.25}$$

A study on the TIMIT corpus has found that unvoiced phonemes accounted for 23.1 % of all the phonemes[33]. The TIMIT corpus consists of 6,300 sentences read by 630 different speakers from 8 major dialect regions in America. Considering this percentage and intervals of silence in the data set a conservative threshold is taken. The objective is

to ensure that the test signal contains minimal unvoiced segments. Loss of some voiced segments is tolerated.

The segments with z_r = 50 (corresponding to 2000 Hz) and $E_{\log}$ > 60% of the median for the whole recording (about 20 s duration) is classified as the voiced segment.

The steps involved in data preparation are summarized below :

1. Up sample the speech signal to 48 kHz.
2. Take the windowed signal (window length of 25 ms.)
3. Find Z_r and E_{log}.
4. Drop the segments with Z_r > 50 and E_{log} < 60 % of the median E_{log} for the recording.
5. After processing all the segments are joined together and is down sampled to 8 kHz after anti- aliasing.

Down sampling smooths out the discontinuities that occur while the segments are joined back together. Figure 5.1 and Figure 5.2 show the speech signals before and after removal of unvoiced and silent segments. The spectrograms of the signals are also shown. In the spectrogram higher energies are represented by darker pixels. It can be observed that in the voiced only signal the segments with higher energy at higher frequencies are removed. Silent segments (low envelope energy) are also removed. Table 5.1 lists the result of voiced-unvoiced- silence segregation for all the signals in the dataset. Only voiced segments are retained.

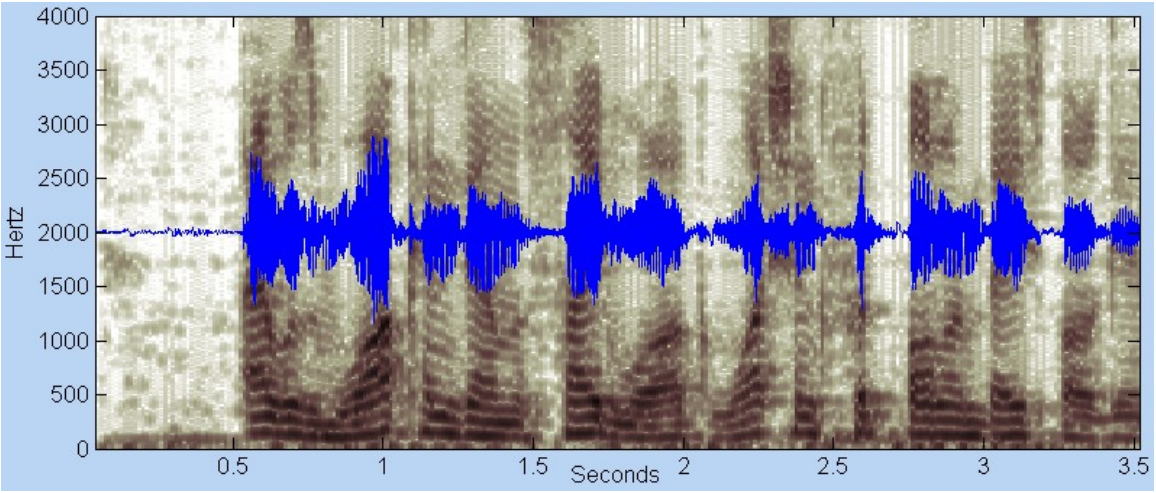

Figure 5.1: Speech signal of a male speaker first 3.5 seconds

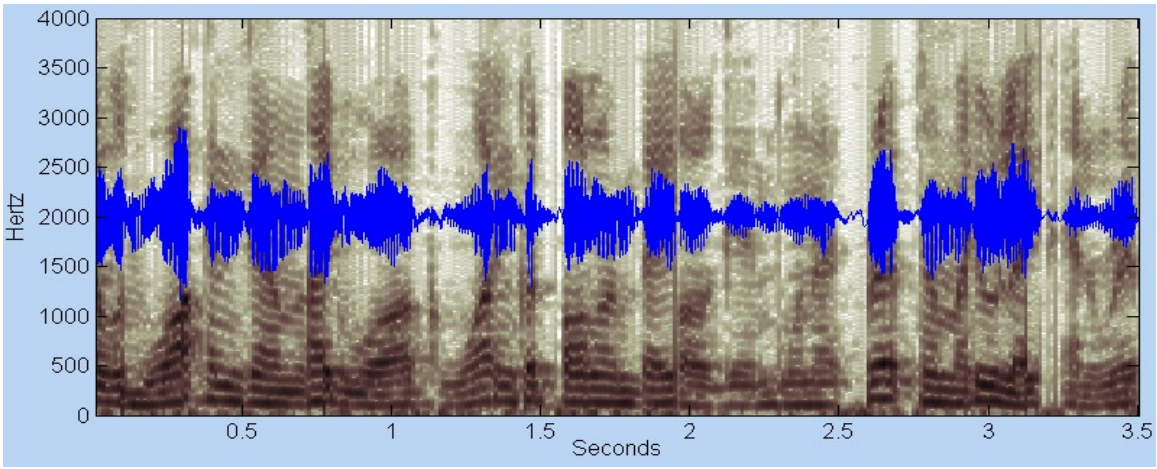

Figure 5.2: Speech signal after unvoiced and silent segments are removed.

The signal shown in 5.1 is the input. When unvoiced segments are dropped the signal is shifted backwards.

Table 5.1: Amount of voiced , unvoiced, silent segments in the dataset

Speaker ID (p)	Speaker	Recording (r)	Voiced	Unvoiced	Silence
1	Male 1	Recording 1	58.45%	27.61%	13.94%
		Recording 2	56.69%	23.25%	20.06%
		Recording 3	52.24%	35.86%	11.90%
2	Male 2	Recording 1	55.77%	28.75%	15.48%
		Recording 2	55.79%	34.15%	10.06%
		Recording 3	50.07%	39.32%	10.61%
3	Male 3	Recording 1	64.03%	22.47%	13.50%
		Recording 2	63.16%	17.52%	19.32%
		Recording 3	63.53%	18.09%	18.38%
4	Male 4	Recording 1	56.69%	27.08%	16.23%
		Recording 2	56.89%	21.19%	21.93%
		Recording 3	49.47%	30.50%	20.03%
5	Male 5	Recording 1	53.31%	29.28%	17.41%
		Recording 2	48.87%	30.26%	20.88%
		Recording 3	51.67%	37.00%	11.33%
6	Female 1	Recording 1	40.72%	42.57%	16.71%
		Recording 2	41.36%	37.83%	20.80%
		Recording 3	39.34%	42.39%	18.27%
7	Female 2	Recording 1	50.54%	37.79%	11.67%
		Recording 2	58.62%	28.82%	12.55%
		Recording 3	39.81%	39.81%	20.37%
8	Female 3	Recording 1	48.48%	17.09%	34.42%
		Recording 2	41.28%	38.08%	20.64%
		Recording 3	41.66%	35.77%	22.57%

5.4. **Feature Analysis**

The voiced speech data created in the previous section is used for feature analysis. The features are found for each recording for every 50 ms window. Features extracted from each recording are then averaged together to obtain a reference template:

$$\Omega_{p,r}[n]=\frac{1}{N_W}\sum_{n_w=1}^{N_W} L_x[n,n_w;p,r] \tag{5.26}$$

where $N_W = 146$ is the total number of 50ms windows present. r = 1,2 ,3 is the recording per speaker, p = 1,2,..,8 represents the speaker (See Table 5.1) . $L_x[n_w;p,r]$ represents the features presented in section 5.2. for the w^{th} segment. And $\Omega_{p,r}$ is the reference template obtained from the r^{th} recording of p^{th}speaker. Subscript x represents any one of the feature presented in section 5.2. or their combination taking two at a time. $\Omega_{p,r}$ is tested for its ability to identify speaker p from a pair of speakers. The test set for $\Omega_{p,r}$, $\mathbb{C}$ is given by the pair :

$$\mathbb{C} = \left\{ L_x[n_w; p, l], L_x[n_w; q, m] \right\} \qquad \forall\, q \neq p, l \neq r \qquad (5.27)$$

The condition $l \neq r$ ensures that the performance measured will be speech independent.

5.4.1. Distance measure and classifier

A minimum distance classifier is used to distinguish between the speakers q and p. The Mahalanobis distance from the reference $\Omega_{p,r}$ to $L_x[n_w; p, l]$ and $L_x[n_w; q, m]$ is obtained as :

$$D_M[n_w; L_{x,p,l}, \Omega_{p,r}] = \sqrt{(L_x[n_w; p, l] - \Omega_{p,r})^T \, \Sigma^{-1}\left(L_x[n_w; p, r]\right) (L_x[n_w; p, l] - \Omega_{p,r})}$$

$$D_M[n_w; L_{x,q,m}, \Omega_{p,r}] = \sqrt{(L_x[n_w; q, m] - \Omega_{p,r})^T \, \Sigma^{-1}\left(L_x[n_w; p, r]\right) (L_x[n_w; q, m] - \Omega_{p,r})} \qquad (5.28)$$

where Σ represents the covariance matrix. The dependency of L on the speaker and the recording is denoted as suffix from now on. The time sample index n is dropped for readability. Mahalanobis distance gives each dimension of the feature vector a weight which is dependent on its variance across time. Higher variance will result in lesser weight.

The speaker is detected as :

$$\hat{p} = argmin_{p/q}\left\{ D_M[n_w; L_{x,p,l}, \Omega_{p,r}], D_M[n_w; L_{x,q,m}, \Omega_{p,r}] \right\} \qquad (5.29)$$

The truth hypothesis can be defined as $H_p : \hat{p} = p$. Probability of true detection for a given speaker and recording is:

$$P(H_p|\Omega_{p,r}, L_x) = \frac{\mathrm{n}(\hat{p} = p)}{\mathrm{n}(\hat{p})} \tag{5.30}$$

where n(.) represents the number of (.). The probability of true detection for a given speaker is given by :

$$P(H_p|\Omega_p, L_x) = \frac{1}{N_r} \sum_{r=1}^{N_r} P(H_p|\Omega_{p,r}) \tag{5.31}$$

Variation in $P(H_p|\Omega_p, L_x)$ across the speakers will give a measure of dependency of the feature on the speaker. The probability of true detection for a feature is given by :

$$P(H_p|L_x) = \frac{1}{N_r N_p} \sum_{1}^{N_p} \sum_{r=1}^{N_r} P(H_p|\Omega_{p,r}) \tag{5.32}$$

The feature or the combination of features which give the highest $P(H_p|L_x)$ is selected to be used for linking the localized streams.

5.4.2. Feature length/dimension

The length/dimension of the multi dimension features namely, *MFCC*, Δ *MFCC* , GFCC Δ *GFCC* and Vocal Tract Impulse Response are empirically estimated. $P(H_p|L_x)$ is determined by increasing the dimension from 1 in steps of 1. The length is not further increased if there is no further significant improvement in the true detection rate. Table 5.2. lists the features and their dimensions.

Table 5.2: Auditory features used and their Dimension

Auditory Features used for ASS	Dimension
Power	1
Pitch	1
Rate	1
MFCC	28
Δ MFCC	23
Vocal tract Impulse Response	12
GFCC	26
Δ GFCC	23

5.5. **Result and Discussion**

Each feature is analyzed for dependence on speaker and gender (same or different). Table 5.3 shows the number of true detections when it is attempted to detect MALE1($p=1$) in all possible dataset as mentioned in Eq.5.27. There are 292 decisions made for each combination of (p,r). Table 5.4 shows the consolidated True Detection Rate (TDR) for all the detections with $p=1$ (Male1). As expected the detection rate when the speakers are of different gender is considerably higher than if they are of same gender. Similar analysis was carried out for all the speakers (p = 1,2, …, 8) and the TDR is listed in Table 5.5.

Table 5.3: TDR using GFCC

	True Detection Rate		
Voice (*q, m*)	$r = 1; l = 2,3$	$r = 2; l = 1,3$	$r = 3; l = 1,2$
MALE2 Recording1	52.40%	44.18%	57.19%
MALE2 Recording2	68.15%	53.77%	65.07%
MALE2 Recording3	62.67	58.56%	65.75%
MALE3 Recording1	74.32%	69.18	76.37%
MALE3 Recording2	64.38%	58.90%	63.70%
MALE3 Recording3	59.25%	50.34%	64.04%
MALE4 Recording1	48.29%	50.34%	64.04%
MALE4 Recording2	48.97%	51.37%	51.71%
MALE4 Recording3	50.34%	56.51%	45.89%
MALE5 Recording1	59.25%	57.88%	60.62%
MALE5 Recording2	69.18%	66.78%	66.10%
MALE5 Recording3	68.49%	59.93%	58.90%
FEMALE1 Recording1	83.90%	82.53%	84.93%
FEMALE1 Recording2	83.22%	79.11%	81.85%
FEMALE1 Recording3	82.53%	75.34%	79.45%
FEMALE2 Recording1	82.53%	76.03%	77.40%
FEMALE2 Recording2	83.22%	77.74%	84.59%
FEMALE2 Recording3	84.93%	76.71%	81.16%
FEMALE3 Recording1	73.29%	59.93%	76.37%
FEMALE3 Recording2	86.30%	73.29%	81.51%
FEMALE3 Recording3	81.51%	75.00%	78.42%

Data shown for MALE 1 (p =1), Number of decisions = 292

Table 5.4: Consolidated TDR for GFCC , MALE1

	$r = 1$	$r = 2$	$r = 3$	Overall
TDR	69.86%	64.60%	68.97%	67.81%
TDR Same gender	60.47%	56.74%	60.22%	
TDR Cross gender	82.38%	75.08%	80.63%	

Table 5.5: TDR with GFCC for all speakers

Speaker	Success Rate
MALE1	67.81%
MALE2	69.51%
MALE3	73.66%
MALE4	82.23%
MALE5	78.26%
FEMALE1	84.16%
FEMALE2	77.94%
FEMALE3	77.39%
Mean	76.37%
Std Deviation	5.73%

Table5.6 lists the TDR for all the features analyzed. It can be observed that GFCC and ΔGFCC performed better for text independent speaker identification when compared to other proposed features. Pitch acted as a good classifier when the speakers are of different genders. The performance of pitch decreased drastically when both speakers were of the same gender.

The features were also analyzed appending two at a time. The order in which they are combined will not affect the performance. Figure 5.3 shows a plot of TDR for all the possible combination of features taking two at a time. The top 10 performing feature combinations are listed in Table 5.7. It can be observed that GFCC combined with other

features outperforms other analyzed features. It is also seen that the GFCC-pitch combined outperforms GFCC alone by only about 1.1%.

Table 5.6: TDR for various features (Tested on voiced segments)

Feature	True Detection Rate (%)	Standard Deviation (w.r.t to speaker, %)	TDR Same gender(%)	TD Cross gender(%)
GFCC	76.37	5.73	67.96	82.42
Δ GFCC	75.67	4.88	68.29	81.27
Pitch	70.01	7.74	58.75	79.78
MFCC	65.42	5.75	60.53	69.30
Vocal Tract Imp. Response	65.35	5.67	63.84	65.92
Δ MFCC	65.31	4.94	60.11	69.64
Center of Mass of (H_v)	53.03	4.61	52.72	54.19
Power	49.38	5.64	49.56	50.68
Rate	49.04	3.05	48.81	50.57

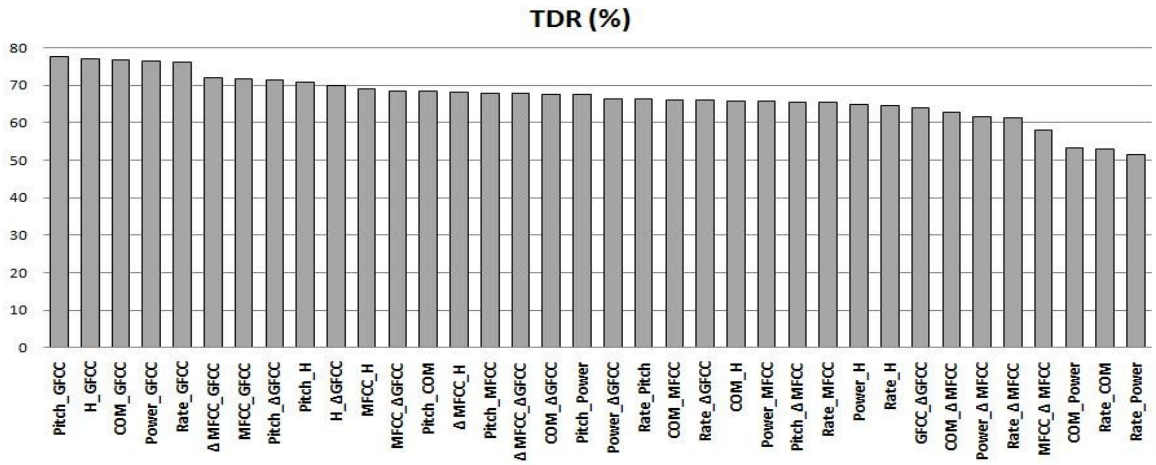

Figure 5.3: TDR for Auditory features taking two at a time (sorted high to low)

Table 5.7: TDR for auditory features taken 2 at a time(Tested on voiced segments)

Rank	Features	TDR	Std. Dev.	TDR Same gender	TDR Cross gender
1	Pitch_GFCC	77.48%	5.90%	68.25%	84.39%
2	H_v_GFCC	77.02%	5.86%	69.28%	82.68%
3	COM_GFCC	76.61%	6.03%	68.45%	82.55%
4	Power_GFCC	76.44%	5.52%	67.99%	82.53%
5	Rate_GFCC	75.99%	5.53%	67.57%	82.09%
6	Δ MFCC_GFCC	72.11%	5.95%	65.34%	77.28%
7	MFCC_GFCC	71.62%	6.45%	64.98%	76.69%
8	Pitch_ΔGFCC	71.40%	5.96%	65.22%	77.64%
9	Pitch_H_v	70.86%	6.24%	65.48%	75.58%
10	H_ΔGFCC	69.90%	4.77%	66.04%	73.58%

5.6. **Conclusion**

This chapter introduced and analyzed various auditory features which can be used for ASS. Clean speech recordings of eight speakers were used to test the features. Only voiced segments of speech were used. It was observed that the combination of GFCC and Pitch gave the best performance of all of them. It gave a TDR of 77.5% while classifying between two speakers. The decision was made using 50ms of clean speech. For linking the localized streams there would be many 50ms. windows available in each stream. This would result in a higher TDR. In the following chapter, the combination GFFC- Pitch is used to link the localized streams obtained in Chapter 4.

Chapter 6. Auditory Features for ASS on Beamformed Signals

6.1. Introduction

ASS using spatial cues resulted in over segmentation and produced spatially localized streams. These streams should be linked across time to represent the same object (human speaker) irrespective of their position. The high level features used in CASA and speaker recognition tasks can be used for this. A few of such features were analyzed in Chapter 5. Combination of GFCC and Pitch was found to be performing better than the other analyzed features(Table 5.6 and Table 5.7). In this chapter these auditory features are extracted after DS beamforming on localized stream locations. Then their ability to classify the streams as speaker1 or speaker2 is tested. The actual positions of the speakers are known, and hence the classifier can be tested for its accuracy.

6.2. Beamforming on Localized Streams

The experiment run in Chapter 4 resulted in a set of positions where source detections were denoted by H_ζ for segment index ζ (Eq.4.9). The set of respective streams associated with each position are obtained with:

$$y_\zeta(t) = \mathrm{B}\left(\boldsymbol{H}_{\boldsymbol{\zeta}}; \boldsymbol{X}_N\right) \tag{6.1}$$

where $\mathrm{B}(\cdot)$ represents DS beamforming(Eq.2.4). $\boldsymbol{X}_N$ is the array of signals at N microphones(16 in this thesis). The beamformer target location is a time varying function and is determined by the elements of H_ζ arranged sequentially in time. The beamforming is carried out every 20ms. Sometimes due to intervals of silence or miss detections the target coordinates may not be available for all the time instants. In that case the most recently available coordinate is used. Table 6.1 shows the first few points for stream $\boldsymbol{H}_1$. The spatial coordinates are estimated every 20ms. But it can be observed that coordinates are not available at 3.48 through 3.56 seconds. Hence during this period the beamformer will target at (2.04, 2.04)m; location estimated at 3.48sec. Similar discontinuities can be observed at many instances.

Table 6.1: Stream 1 (H_1 , Tracking information (first few points).

Time (s)	*x* coordinate (m)	*y* coordinate (m)
3.46	2.04	2.04
3.48 – 3.56	*Miss detections*	
3.58	2.04	2.04
3.60	2.04	2.04
3.62	2.04	2.04
3.78	2.04	2.08
3.80	2.04	2.04
3.82 – 3.88	*Miss detections*	
3.90	2.04	2.04
4.14	2.04	2.04
4.16	2.04	2.04
4.18	2.04	2.04
4.20	2.04	2.04
4.22	2.04	2.04
4.24	2.04	2.04

6.3. **Binary Least Mahalanobis Distance Classifier**

It is attempted to classify the detections into either speaker1 or speaker2. Let $\boldsymbol{S_p}$, $\boldsymbol{S_q}$ represent the set of all streams belonging to speaker1 and speaker2 respectively. Also let the auditory feature vector representing each stream $y_\zeta(t)$ be represented by $L_\zeta[n_w]$. Then by Eq.5.28 the Mahalanobis distance to a reference signal can be calculated as:

$$D_M\left[L_\zeta[n_w];\Omega_{\zeta_{ref}}\right]=\sqrt{(L_\zeta[n_w]-\Omega_{\zeta_{ref}})^T\,\Sigma^{-1}(L_{\zeta_{ref}}[n_w])\,(L_\zeta[n_w]-\Omega_{\zeta_{ref}})} \tag{6.2}$$

where $\Omega_{\zeta_{ref}}$ is the reference signal and is obtained by :

$$\Omega_{\zeta_{ref}}=\frac{1}{N_W}\sum_{n_w=1}^{N_W} L_{\zeta_{ref}}[n_w] \tag{6.3}$$

where N_W is the number of 50ms segments present in the reference stream $y(t)_{\zeta_{ref}}$ and n_w indexes the 50ms non-overlapping analyzing windows. The candidate stream is compared with the reference signal and a preliminary decision is made every 50ms.

One localized stream is selected from each set $\boldsymbol{S_p}$, $\boldsymbol{S_q}$ and the reference feature vector represented by Ω_p and Ω_q are formed by Eq.6.3. Let $y_p(t)$ and $y_q(t)$ represent the streams of speaker1 and speaker2. Then the preliminary decision is made by :

$$\frac{D_M\left[L_p, \Omega_p\right]}{D_M\left[L_q, \Omega_P\right]} \underset{\hat{p}=q}{\overset{\hat{p}=p}{\lessgtr}} 1 \text{ and } \frac{D_M[L_p, \Omega_q]}{D_M[L_q, \Omega_q]} \underset{\hat{q}=q}{\overset{\hat{q}=p}{\lessgtr}} 1 \tag{6.4}$$

Implicit dependency on time index n_w is dropped for readability. Ideally $\hat{p}=p \Rightarrow \hat{q} \neq p$ and vice versa. But in reality there can be conflicts. A conflict is said to occur when $\hat{p}=\hat{q}$. The conflict in which p is assigned to both $\hat{p}$ and $\hat{q}$ is resolved by :

$$\frac{D_M\left[L_p, \Omega_p\right]}{D_M\left[L_p, \Omega_q\right]} \underset{\hat{q}=p}{\overset{\hat{p}=p}{\lessgtr}} 1 \tag{6.5}$$

If the conflict is with the assignment of *q,* then *p* is replaced by *q* in Eq.6.5.

If *l* preliminary decisions are made, then the final classification is based on the "majority criterion" rule. i.e :

$$\tilde{p}=\begin{Bmatrix} p\,;H_{p/q} \geq 0.50 \\ q\,;H_{p/q} < 0.50 \end{Bmatrix} \qquad \text{where } H_{p/q}=\frac{\mathrm{n}(\hat{p}=p)}{l} \tag{6.6}$$

True Detection Rate is defined as :

$$D_t \overset{\text{def}}{=} \frac{\mathrm{n}(\tilde{p}=p)}{\mathrm{n}(\tilde{p}=p)+\mathrm{n}(\tilde{p}=q)} \tag{6.7}$$

6.4. **Performance Analysis**

6.4.1. The feature vector

From the analysis in Chapter 5. GFCC-Pitch was identified to be the best feature vector among the tested ones. GFCC-Pitch is used here to classify the localized streams as belonging to one speaker or the other. Since GFCC is a multi dimensional feature, its

length is varied from 1 to 26 and the performance is analyzed. The feature vector is obtained by:

$$L = \begin{bmatrix} L_p & L_{G1} & L_{G2} & \dots & L_{GN} \end{bmatrix} \tag{6.8}$$

where L_p is the pitch extracted and $L_{G,N}$ is the N dimensional GFCC. The total feature length is N+1.

6.4.2. Test Data

By comparing the spatial coordinates of H_ζ and the predefined speaker locations the streams corresponding to each speaker in the experiment in Chapter 4 are identified manually. The set of streams belonging to the same speaker is given by (Table 4.4) :

$$\begin{aligned} \boldsymbol{S_p} &= \left\{ y_\zeta(t) : \zeta = 1,2,3,9,10,11,13,15,16,19,20,25,27,29,30 \right\} \\ \boldsymbol{S_q} &= \left\{ y_\zeta(t) : \zeta = 4,5,6,8,14,18,21,23,24,26,28 \right\} \end{aligned} \tag{6.9}$$

One stream each from $\boldsymbol{S_p}$ and $\boldsymbol{S_q}$ is chosen as the reference. The test set is given by the pair:

$$\mathbb{C} = \left\{ L_{p,T}, L_{q,T} \right\}; \text{for } T = 1,2,\dots,6 \tag{6.10}$$

where T represents the Time ID (Table 4.4). Only streams which intersect in time (same Time ID) are paired together for testing. Table 6.2. shows the data set for T = 1. Similar datasets are made for all values of T. $\zeta = 20$ is chosen as the reference for $\boldsymbol{S_p}$ and $\zeta = 18$ is chosen as the reference for $\boldsymbol{S_q}$. The references were empirically chosen with the restriction that they have the same Time ID.

Table 6.2: Test Data Set for Time ID = 1

Time ID	ζ for S_p	ζ for S_q	N
1	1	4	44
		5	31
		6	49
	2	4	44
		5	31
		6	50
	3	4	44
		5	31
		6	50

$D_M\left[L_p, \Omega_p\right]$ and $D_M\left[L_p, \Omega_q\right]$ are computed and preliminary decisions are made l times where l is given by :

$$l = \min\left\{N_p, N_q\right\};$$
$$N_p = \left\lceil \frac{T_p}{50\text{ms.}} \right\rceil; \quad N_q = \left\lceil \frac{T_q}{50\text{ms}} \right\rceil \tag{6.11}$$

where T_p and T_q are the length of streams $y_p(t)$ and $y_q(t)$ respectively.

6.4.3. Results

At Time ID = 1 there are 9 test cases possible (Table 6.2). Figure 6.1 shows $H_{p/q}$ (in percentage) for all the 9 instances. In all cases $H_{p/q} > 50\%$ which implies that TDR, $D_t = 100\%$ for Time ID,T = 1. The same test is carried out for T = 1, 2, …, 6. and N = 1,2, …, 26. The result is shown in Figure 6.2. The intensity is proportional to D_t. It is observed that the best result is achieved when N = 21 (feature length of 22). At T = 4 TDR of 100% is obtained for all feature length. This is because the feature vectors also have the same time ID.

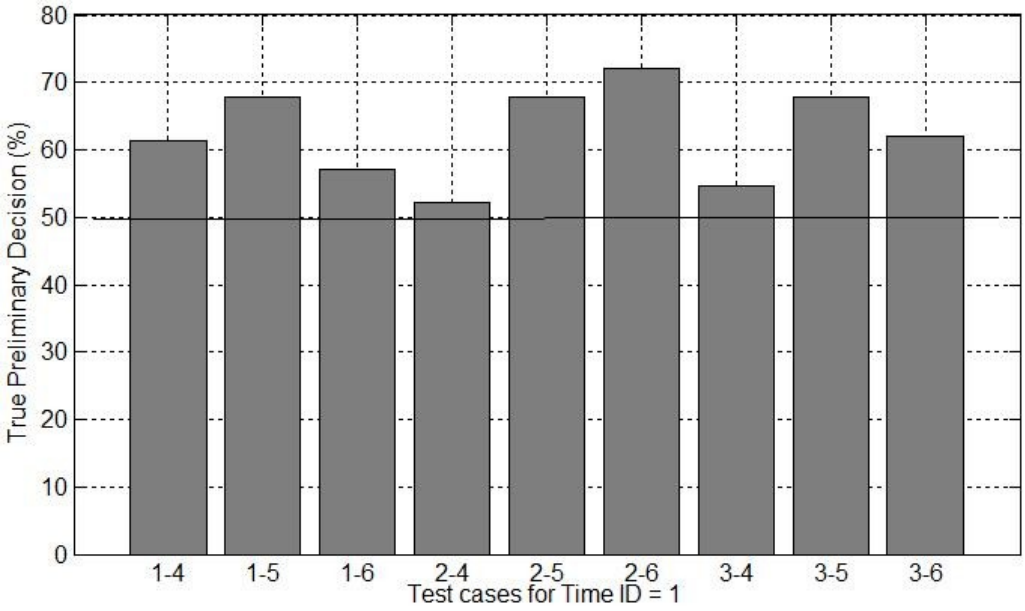

Figure 6.1: Percentage of correct preliminary decisions for the binary classifier. The horizontal line (at 50%) marks the boundary for final decision. Feature length (N + 1) = 22.

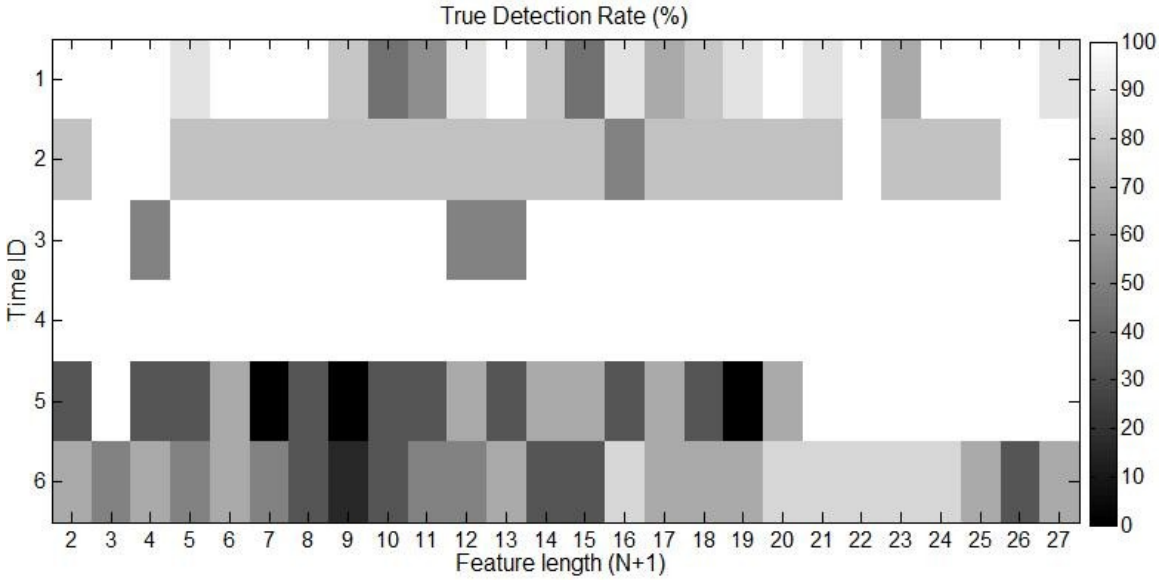

Figure 6.2: TDR for the binary classifier after applying majority criterion.

Table 6.3 lists the performance of the classifier. N = 21 is used as it gave the best performance. A 65.17% of preliminary decisions made were true.

Table 6.3: True Detection Rate for Binary classifier; N =21

Time ID	Distance between Speakers (m)	Number of Preliminary Decision	True Preliminary Decision(%)	Number of final Decision	True Detection Rate (%)
1	1.589	374	62.03	9	100
2	1.534	88	69.32	4	100
3	1.404	85	61.18	2	100
4	1.683	101	95.05	2	100
5	1.699	140	61.43	3	100
6	1.058	168	57.14	6	83.33
Total	-	956	65.17	26	96.15

Figure 6.3 shows TDR of preliminary and final decisions as a function of the feature length.

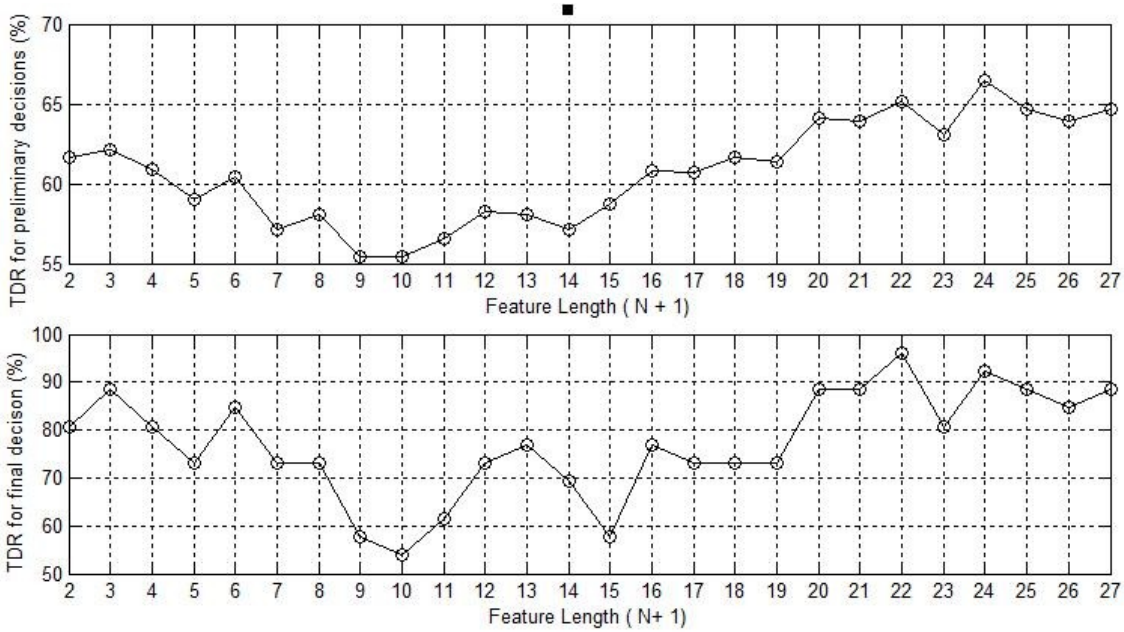

Figure 6.3: TDR as a function of feature length.

6.5. **Conclusion**

In this chapter a binary classifier which classifies the localized streams into two streams (representing two speakers) was designed and tested. The results obtained demonstrate that auditory features extracted from the beamformed signal can be used to link the speaker detections across time. Out of 26 decisions made 25 of them are correct. i.e. a final TDR of 96.15 % is achieved. With around 2 seconds (40 preliminary decisions) of localized streams available the streams can be linked to the correct speaker with over 90% accuracy.

The performance of the classifier for varying length of GFCC is also studied. Apart from the unstable behavior for N = 1,2, ..., 7 the general trend is that the TDR improved with increasing length of GFCC. The TDR seems to oscillate around 90 % for $N > 19$ (feature length of 20, Figure 6.3). Any further increase in length may not bring any significant improvement. The fluctuation in TDR is due to the small population size. TDR for preliminary decision is a better indicator as the population size is high (956).

Chapter 7. Conclusion and Future Work

7.1. Overview

This thesis aimed at extracting the streams representing distinct sources in the audio scene. Specifically the case of two simultaneous talkers was considered. A microphone array was used for localizing the sound sources and then to beamform on them. Spatial and temporal thresholds were applied to obtain localized streams. The system using just these thresholds was unable to track the speaker when he/she moved to a new position while remaining silent. This necessitated the use of auditory features for merging the spatially localized streams. Auditory features namely GFCC, pitch, MFCC, vocal tract impulse response, loudness and rate of speech envelope were analyzed using clean speech recordings. Pitch appended with GFCC outperformed other examined features. Hence it was used for audio scene segmentation and the result is noted.

7.2. Conclusion

The following was demonstrated in this work :

1. Gammatone Frequency Cepstral Coefficients (GFCC) along with pitch of the speakers gave an accuracy of 96.2 % in separating the streams belonging to two simultaneous speakers. This demonstrates the viability of using them as features for Audio Scene Segmentation.
2. Feature length for optimum performance is estimated as 22; pitch appended with the 21 point GFCC.
3. In the clean speech analysis done in Chapter 5, GFCC gave a true detection rate of 76.4% compared to MFCC (65.42 %). The limited study done presents a case for using GFCC as a feature for automatic speaker recognition.

7.3. **Future Work**

This thesis opens up a few issues which need further analysis. First among them would be to evaluate the presented system as a function of the beamformer performance. It is obvious that the features extracted will be more reliable with higher beamformer performance. A fall in TDR from 77.48 % for clean speech to 65.15% in the case of simultaneous speakers (refer Table 5.7 and Table 6.3) demonstrates this. A metric for beamformer performance evaluation may be required.

The spatial separation between the speakers and their position in the beamfield of the array may have an effect on the performance. Future experiments will have to take this into account. The speakers can be placed in accordance with the beam pattern (in the main lobe area, nulls etc).

The tests were conducted only for a two speaker scenario. For a more generic solution a threshold for the distance from the reference feature must be obtained. With a threshold value setup, the feature vector extracted from the any one of the localized streams can be used as reference and audio scene segmentation can be performed. Any unassigned localized stream can be iteratively merged to one of the final streams. This will make the algorithm independent of the number of speakers. This calls for threshold estimation using a larger database.

Due to the complexity of the analysis only one experimental set up was used for performing audio scene segmentation. i.e. array geometry and the pair of speakers remained the same for the whole analysis. The performance analysis for different array geometries may be carried out. Also using different combinations of speakers (both male and female) will help in generalizing the results further.

REFERENCES

[1] Albert Bergman, *Auditory Scene Analysis*, MA, USA: MIT press,1990.

[2] Martin Cook & Daniel P W Ellis, "*The Auditory Organization of Speech and other Sources in Listeners Computational Models", Speech Communication* , vol. 35, no. 3-4, pp. 141-171, Oct. 2001.

[3] Wang D & Brown GJ (Eds.)*Computational Auditory Scene Analysis: Principles, Algorithms and Applications* , Wiley/IEEE press, NY, USA, 2006.

[4] Kathy Melih & Ruben Gonzalez,*Harmonic grouping for Computational Auditory Scene Analysis*, 8th International Symposium on Signal Processing and its Applications, 2005, Sydney, Australia.

[5] Heinz Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, Heidelberg, NY, USA: Springer,2007.

[6] Kevin D Donohue, Jens Hannemann & Henry G Dietz, "*Performance of Phase Transform for Detecting Sound Sources with Microphone Arrays in Reverberant and Noisy Environments", Signal Processing* , vol. 87, no. 1, pp. 1677 - 1691, Jan. 2007.

[7] Anand Ramamurthy, Harikrishnan Unnikrishnan & Kevin D Donohue, *Experimental Performance Analysis of Sound Source Detection with SRP PHAT-* β IEEE SoutheatCon,2009, Atlanta, GA, USA.

[8] M Coen, *Design Principles for Intelligent Environments*. proceedings of Fifth National Conference on Artificial Intelligence,1998, Madison, WI, USA.

[9] J Benesty, J Chen & Y Huang, *Conventional Beamforming Techniques*, Microphone Array Signal Processing, , Berlin, Germany: Springer,2008, pp. 39 - 64.

[10] Simon Haykin & J Justice, *Array Signal Processing*, Englewood Cliffs, NJ, USA: Prentice Hall,1985.

[11] O Hosyuma & A Sugiyama, *Robust Adaptive Beamforming*, Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition, Brandstein M & Ward D (Eds.), Berlin: Springer,2007, pp. 149 - 178.

[12] Anand Ramamurthy, *Experimental Evaluation of Modified Phase Transform for Sound Source Detection*, Masters Thesis, University of Kentucky, Lexington, KY, USA, 2007.

[13] Kevin D Donohue, Kevin S McReynolds & Anand Ramamurthy, *Sound Source Detection Threshold Estimation using Negative Coherent Power*. IEEE SouteastCon,2008, Hunstville, Alabama, USA.

[14] Kevin D Donohue, Sayed M SaghaianNejadEsfahani & Jingjing Yu,*Constant False Alarm Rate Sound Source Detection with Distributed Microphones*, IEEE Transactions on Signal Processing, to be published.

[15] Jacek Dmochowski, Jacob Benesty & Sofiene Affes, "*On Spatial Aliasing in Microphone Arrays", IEEE Transactions on Signal Processing* , vol. 57, no. 4, pp. 1383-1395, Apr. 2009.

[16] Michael L Seltzer & Bhiksha Raj, *Calibration of Microphone Arrays for improved Speech Recognition*. EUROSPEECH,2001, Denmark.

[17] Behringer©, "*Measurement Microphone ECM8000"*, Specification sheet, Bheringer Spezielle Studiotechnik GmbH, 2000.

[18] Auralex Acoustics, Inc., *MAX-Wall Acoustic Panels*, http://www.auralex.com/testdata/, 2009.

[19] Avid Thecnology, Inc., *Delta 1010™ Digital Recording System*, www.m-audio.com/products/en_us/delta1010.html, 2009.

[20] Avid Technology, Inc., *Audio Buddy™*, www.m-audio.com/products/en_us/AudioBuddy.html, 2009.

[21] Paul Davis, *Jack Connection Kit*, jackaudio.org, 2006.

[22] Ben Gold & Nelson Morgan, *Speech and Audio Signal Processing - Processing and Perception of Speech and Music*, NY, USA: John Wiley & Sons Inc,1999.

[23] JCR Licklider, "*Periodicity Pitch and Place Pitch", Journal of Acoustic Society of America* , vol. 26, no. 5, p. 945, Sep. 1954.

[24] S Grossberg, *Pitch-based Streaming in Auditory Perception*, Musical Networks: Parallel Distributed Perception and Performance, Griffith N & Todd P (Eds.), Cambridege, MA, USA: MIT Press,1996, pp. 117-140.

[25] Ray Meddis & Michael J Hewitt, "*Modeling the identification of Concurrent Vowels With Different Fundamental Frequencies", Journal of Acoustic Society of America* , vol. 91, no. 1, pp. 233 - 245, Jan. 1992.

[26] Matti Karjalainen & Tero Tolonen,*Multi-Pitch and Periodicity Analysis Model for Sound Separation and Auditory Scene Analysis*, .

[27] Tim R Black & K D Donohue, *Pitch Determination of Music Signals Using the Generalized Spectrum*. IEEE SoutheastCon,2000, Nashville, TN, USA.

[28] Yang Shao & DeLiang Wang, *Robust Speaker Identification using Auditory Features and Computaitonal Auditory Scene Analysis*. ICASSP,2008, Las Vegas, Nevada, USA.

[29] Joseph P Campbell, *Speaker Recognition: A Tutorial*. Proceedings of the IEEE,1997, USA.

[30] Fang Zheng, Guoliang Zhang & Zhanjiang Song, "*Comparison of Different Implementations of MFCC", Journal of Computer Science and Technology* , vol. 16, no. 6, pp. 582 - 589, Nov. 2001.

[31] Malcom Slaney, "*An Efficeint Implementation of the Patterson-Holdsworth Auditory Filter Bank",* Technical Report, Apple Computer, Inc., 1993.

[32] Kevin D Donohue, *Single Speaker Recording*, www.engr.uky.edu/~donohue, 2009.

[33] D L Wang & G Hu, *Unvoiced Speech Segregation*. IEEE ICASSP,2006, Toulouse, France.

VITA

Harikrishnan Unnikrishnan was born on September 3, 1984 in Palakkad, Kerala, India. The author received his Bachelor of Technology(B. Tech.) degree in Electronics and Communication from Mahatma Gandhi University, Kerala, India in the year 2006. He has worked as a software engineer in Hexaware Technologies, Chennai, India before enrolling for Masters program in Electrical Engineering at University of Kentucky, Lexington. He has been working at Center for Visualization and Virtual Environments as a Research Assistant for Dr. Kevin D. Donohue since April 2008. He is a student member of IEEE since 2003. The author has received a certificate for appreciation of "outstanding volunteerism" from IEEE, Kerala section in the year 2005.